HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2021 February 26.

Published in final edited form as:

Nature. 2020 September ; 585(7823): 124–128. doi:10.1038/s41586-020-2638-5.

Functionally uncoupled transcription-translation in *Bacillus subtilis*

Grace E Johnson[1,†], **Jean-Benoît Lalanne**[1,2,†], **Michelle L Peters**[1], **Gene-Wei Li**[1,*]

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA 02139

[2]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA 02139

Abstract

Coupled transcription and translation is considered a defining feature of bacterial gene expression[1,2]. The pioneering ribosome can both physically associate and kinetically coordinate with the RNA polymerase (RNAP)[3-11], forming a signal-integration hub for co-transcriptional regulation that includes translation-based attenuation[12,13] and RNA quality control[2]. However, whether transcription-translation coupling – together with its broad functional consequences – is indeed a fundamental characteristic outside the well-studied *Escherichia coli* remains unresolved. Here we show that RNAPs outpace pioneering ribosomes in the Gram-positive model bacterium *Bacillus subtilis*, and that this 'runaway transcription' creates alternative rules for both global RNA surveillance and translational control of nascent RNA. In particular, uncoupled RNAPs in *B. subtilis* explain a diminished role of Rho-dependent transcription termination, as well as the prevalence of mRNA leaders that utilize riboswitches and RNA-binding proteins. More broadly, we identified widespread genomic signatures of runaway transcription in distinct phyla across the bacterial domain of life. Our results demonstrate that coupled RNAP-ribosome movement is not a general hallmark of bacteria. Instead, translation-coupled transcription and runaway transcription constitute two principal modes of gene expression that determine genome-specific regulatory mechanisms in prokaryotes.

[*]To whom correspondence should be addressed. gwli@mit.edu.
[†]These authors contributed equally to this work.

Author contributions: G.E.J., J.-B.L. and G.-W.L. designed experiments; G.E.J. performed induction kinetic experiments, performed Rho and polarity experiments, and analysed sequence features of Rho target RNAs; J.-B.L performed ORF extension experiments, identified intrinsic terminators from Rend-seq data, analysed nested antisense RNAs and expressed pseudogenes, and wrote the phylogenomic bioinformatic terminator identification pipeline; M.L.P. performed induction kinetic experiments in knockout backgrounds. G.E.J., J.-B.L. and G.-W.L. wrote the manuscript.

Competing interests: The authors declare no competing interest.

Supplementary Information is available for this paper.

Data availability statement:
All data generated and analysed during this study are included in this published article (and its supplementary information and source data files). The high-throughput sequencing datasets analysed during the current study are available from the Gene Expression Omnibus repository with accession numbers: GSE53767, GSE95211, GSE108295 (see section "High-throughput expression datasets used (ribosome profiling, Rend-seq)" for details). Uncropped gel source data for Northern blot can be found in Supplementary Figure 1.

Code availability statement: Scripts for terminator identification have been deposited to GitHub (https://github.com/jblalanne/intrinsic_trx_terminator_identifier). Core Rend-seq analysis scripts used can be found on Github (https://github.com/jblalanne/Rend_seq_core_scripts). Other custom scripts used for data analysis are available upon request.

Messenger RNA transcription in *E. coli* is accompanied by a closely trailing ribosome, whose ability to modulate the fate of the transcribing RNAP establishes a key paradigm for bacterial gene regulation. At operon leaders, ribosome pausing provides a signal for transcriptional attenuators, such as those of the *trp* and *his* biosynthetic operons[13,14]. Within coding regions, the proximity between RNAPs and ribosomes further enables the termination factor Rho to selectively abrogate transcription if a premature stop codon is present, thereby initiating a surveillance mechanism analogous to nonsense mediated decay in eukaryotes[2,15]. Rho also suppresses antisense transcription by pervasively targeting RNAPs that lack an accompanying ribosome[16]. As such, tight transcription-translation coupling is a fundamental feature of the central dogma in *E. coli* and related species.

Yet, the interplay between transcription and translation remains largely unexplored in other species, notably in the extensively studied Gram-positive bacterium *B. subtilis*. Previous work has shown that *B. subtilis* and *E. coli* often have strikingly divergent regulatory mechanisms. For example, although many *B. subtilis* operons are regulated via transcriptional attenuators similar to their *E. coli* counterparts, the underlying mechanisms primarily rely on riboswitches or RNA-binding proteins, not a nearby ribosome[14,17-19]. In addition, factors that monitor translation-coupled transcription in *E. coli* — Rho and its adaptor NusG — are dispensable in *B. subtilis* with only mild null phenotypes[20]. Furthermore, nonsense-mediated polarity, i.e., Rho-dependent transcription termination of mRNAs containing premature stop codons, is thought to be rare in *B. subtilis*[21,22]. We demonstrate in this work that these intriguing differences stem from genome-wide ‘runaway transcription’, whereby RNAPs transcribe far ahead of trailing ribosomes.

RNAP outpaces pioneering ribosome

To determine whether RNAPs and pioneering ribosomes are kinetically coupled, we first used classic assays based on IPTG-inducible *lacZ* adapted from *E. coli* (1020 aa) (Supplementary Data 1) to measure the times of first appearance for its mRNA and protein (Fig. 1a)[4,8,23-25]. At each time point after induction, an aliquot of cell culture was added to a stop solution (phenol-ethanol for mRNA or chloramphenicol and erythromycin (for *B. subtilis*) for protein). Because protein folding takes place after cell harvesting, it does not prohibit the measurement of the much faster translation kinetics[4,8,23,25]. If the pioneering ribosome were closely following the RNAP, as previously demonstrated and independently confirmed here for *E. coli* (Fig. 1b)[4,8], the first full-length mRNA product would appear at the same time (τ_{TX}) as the first protein product (τ_{TL}). Surprisingly, *B. subtilis* exhibited a substantial delay between the protein and mRNA signals during rapid growth (τ_{TL} - τ_{TX}=40±4 s at a growth rate of 2 h^{-1}) (Fig. 1b). Whereas the translation time is similar to what was measured in *E. coli* (τ_{TL}=77±2 s vs τ_{TL}=78±1 s in *E. coli*), the transcription time is much shorter (τ_{TX}=37±3 s vs τ_{TX}=79±1 s in *E. coli*). A time delay was also observed during slow growth (τ_{TL} -τ_{TX}=32±1 s at a growth rate of 0.65 h^{-1}) (Extended Data Fig. 1). Thus, the RNAP ‘runs away’ from the pioneering ribosome and the two do not reach the end of *lacZ* at the same time.

Runaway transcription also occurs for endogenous *B. subtilis* genes, in addition to the exogenous *lacZ*. To measure their products’ first appearance times, we adapted an α-

complementation-based strategy by appending the C-terminus of endogenous genes with a short region encoding LacZα (100 aa)(Fig. 1c)[25]. The fusion gene is similarly placed under an IPTG-inducible promoter, while keeping the native *B. subtilis* ribosome binding site to maintain translation efficiency. After inhibiting protein synthesis, protein levels were estimated by allowing fusion proteins to complement the beta-galactosidase activity of pre-expressed LacZω post-harvesting (Fig. 1c). The translation time of the full-length reporters (τ_{TL}) is reproducibly longer than transcription time (τ_{TX}): τ_{TL} - τ_{TX} =34±2 s for the *pycA-lacZ*α fusion (1255 aa) and 16±0.5 s for the shorter *tkt-lacZ*α fusion (774 aa) (Fig. 1d). By the time the protein signal starts to accumulate, mRNA levels have already increased by 5- to 20-fold (Fig. 1d). Different methods for cell harvesting and translation inhibition reproducibly yielded the same time delay (Extended Data Fig. 2b). These results show that pioneering ribosomes lag far behind each RNAP, potentially due to mismatches in their respective elongation rates.

To estimate the elongation rates of transcription and translation, we measured τ_{TX} and τ_{TL} for a truncated *pycA-lacZ*α construct (255 aa) under the same promoter and native 5' UTR. Based on the length of truncation (ΔL) and the reduction in first appearance times ($\Delta\tau_{TX}$ and $\Delta\tau_{TL}$), we obtained average transcription and translation elongation rates of 73±2 nt/s and 47±3 nt/s, respectively (assuming the same transcription and translation initiation rates between long and truncated constructs, Fig. 1e and Extended Data Fig. 2c). The transcription elongation rate for this region is nearly twice as fast as mRNA transcription elongation in *E. coli*[4]. By contrast, the translation elongation rate is similar to previous estimates in *E. coli*[4,24,25]. Together, these results show that RNAPs consistently outpace pioneering ribosomes in *B. subtilis*. The mismatch in elongation rates (Δv=26±4 nt/s) create gaps between RNAPs and the trailing ribosomes along nascent mRNAs, reaching ≈360 nt after RNAP has transcribed a gene of ≈1 kb (or a larger gap if translation initiation is slow) (Fig. 1f, Supplemental Discussion).

The difference in transcription elongation rates between *E. coli* and *B. subtilis* could arise from a variety of sources. Because the *lacZ* transcription time is different between these species, the differential speed is unlikely to be due to their nucleic acid sequence or secondary structures (Fig. 1b). We also ruled out contributions of non-essential RNAP subunits and transcription elongation factors, as *B. subtilis* cells lacking either the ω subunit, the ε subunit, GreA, or NusG do not exhibit substantially longer transcription times (Extended Data Fig. 3, Supplemental Discussion). It remains possible that the accelerated speed in *B. subtilis* is driven by differences in the RNAP pause signals[26] or other components of the RNAP[10,11] (Extended Data Fig. 4, Supplementary Discussion). Irrespective of the underlying driver, runaway transcription and the long ribosome-free nascent mRNA indicate that the *B. subtilis* genome must have different rules for co-transcriptional regulation than *E. coli*.

RNAP is insensitive to translation

A major prediction of runaway transcription is that *B. subtilis* RNAPs should be insensitive to translation, which stands in contrast to the pervasive usage of translation-controlled transcription termination in *E. coli*. Consistent with this prediction, we found that intrinsic

transcription terminators in *B. subtilis* are effective even when the entire terminator hairpin is translated. When we forced translation through a strong intrinsic terminator downstream of a highly translated gene (*pupG*) by replacing its stop codon (Fig. 2a and Extended Data Fig. 5a), transcription readthrough remained undetectable (Fig. 2b). By contrast, the same ORF-extended construct in *E. coli* shows completely abrogated terminator activity, which is consistent with the current paradigm that the closely trailing ribosome blocks terminator hairpin formation (Fig. 2c)[12,27,28]. Thus, unlike *E. coli*, the pioneering ribosomes in *B. subtilis* do not strongly modulate intrinsic transcription termination signals (except in some circumstances[29-31]). This fundamental difference may contribute to the divergent regulatory mechanisms observed between these species in two major ways. First, it enables intrinsic terminators to generate functional nonstop mRNAs for highly translated genes in *Bacilli*[32], such as the mRNA for an alternative ribosome rescue factor whose analogue in *E. coli* is generated post-transcriptionally by RNase III cleavage[33,34]. Second, the lack of coupling likely leads to avoidance of ribosome-controlled transcriptional attenuators and instead favors riboswitch- or protein-based mRNA leaders that are widely observed in *B. subtilis*[14,17-19].

The large RNAP-ribosome spacing in *B. subtilis* also profoundly impacts the position of intrinsic transcription terminators at the end of operons. Using end-enriched RNA-seq (Rend-seq, see SI discussion for a description) to map active intrinsic terminators[35], we found that many terminator stem-loops directly overlap with stop codons of the last genes of operons (107/1228 with 'stop-to-stem distance' ≤0 nt, Fig. 2d), a configuration that would cause antitermination in *E. coli* by the pioneering ribosomes[12]. Furthermore, the large majority (72%) of intrinsic terminators in *B. subtilis* are positioned within half a ribosome footprint downstream of the stop codon (stop-to-stem distance ≤12 nt, Fig. 2e, Supplementary Data 2). In *E. coli*, only 24% of intrinsic terminator stems are within 12 nt of the stop codon, and these have significantly reduced termination activity compared to other terminators ($p<10^{-3}$, Extended Data Fig. 5b-e). By contrast, the gene-proximal terminators in *B. subtilis* do not show significantly weaker termination ($p>0.3$, Extended Data Fig. 5f-i), indicating a lack of translational interference on terminator hairpin formation. These results further demonstrate that most operons in *B. subtilis* are transcribed without a closely trailing ribosome.

Rho-termination has alternative roles

This pervasive runaway transcription indicates that the physiological roles of Rho in *B. subtilis* should not involve surveillance of nascent mRNA translation, as this would render most transcription events unproductive. Consequently, *B. subtilis* may lack an important mechanism for removing erroneously transcribed mRNAs that have premature stop codons. Indeed, introducing nonsense mutations early in the *pycA* or *tkt* genes does not lead to substantial decreases in mRNA levels (Fig. 3a), which is consistent with previous reports of weak nonsense polarity[20,21]. Furthermore, the levels of both wildtype and nonsense mRNAs remain unchanged when *rho* is deleted, indicating that Rho does not terminate them despite a long stretch of untranslated RNA (Fig. 3a). Transcriptome analysis for cells lacking Rho also showed limited changes in global mRNA expression despite the pervasive uncoupling as indicated by gene-proximal terminators (Fig. 3b). Operons unaffected by Rho include

many that naturally contain long stretches of untranslated regions, such as nested antisense RNAs (n=29/35 unaffected in *B. subtilis*; 2/3 in *E. coli,* Fig. 3b, Extended Data Fig. 6, Supplementary Data 3) and pseudogenes with interrupted reading frames (8/8 without polarity in *B. subtilis*; 2/8 in *E. coli*, Fig. 3b, Extended Data Fig. 7-8, Supplementary Data 3). Hence, *B. subtilis* does not appear to use Rho to initiate the removal of aberrant nonsense mRNAs and lacks a type of RNA quality control that is thought to be universally present in all domains of life[36].

How, then, does Rho target specific RNAs for termination in *B. subtilis*? Although Rho does not affect most mRNA transcription, it selectively terminates several operons and removes many antisense RNAs[37]. Without being influenced by translation, Rho-termination may be solely dependent on *cis*-encoded elements. Indeed, forcing translation of a 678 nt Rho-terminated antisense RNA (antisense to *cssS*, designated *cssS*^{AS}, Fig. 3c) by inserting it in the reading frame of the *tkt-lacZα* mRNA reporter (with the 7 *cssS*^{AS} stop codons in the *tkt* reading frame replaced by sense codons, Supplementary Data 1) did not abrogate Rho activity (Fig. 3d). The >100× decrease in downstream mRNA level was restored by deleting *rho* (Fig. 3d). Rho-termination is specified by a 339 nt window within *cssS*^{AS} with a high C-to-G ratio, a feature of Rho utilization (*rut*) sites in *E. coli* (Fig. 3d,e)[16]. Thus, the sequence of this antisense RNA alone was sufficient to promote Rho termination, independent of genomic location or translation. More broadly, we found that C-to-G ratio distinguishes Rho-terminated mRNAs and asRNAs from those not terminated by Rho (Extended Data Fig. 9). Together, these results support our prediction that Rho-termination is mechanistically independent of translation in *B. subtilis*. Instead, we propose that Rho is guided by strategically placed *cis* elements, including regions of high C-to-G ratios, in the *B. subtilis* genome to prevent both pervasive transcription in the antisense direction and premature termination in the sense direction.

Runaway transcription across eubacteria

Lastly, we evaluated the prevalence of runaway transcription in other bacterial species. Using the short distance between intrinsic terminator hairpins and the preceding stop codons as a conservative signature for lack of kinetic coupling, we systematically annotated the positions of intrinsic terminators for sequenced bacterial genomes by developing a computational classifier with a low false discovery rate (≤1%, Extended Data Fig. 10a-c, Supplementary Data 4-5). The phylogenetic tree bifurcates into phyla with either many or few short-distance terminators (stop-to-stem ≤12 nt) (Fig. 4). Over half of Firmicutes (182/358 analyzed), which includes *B. subtilis* and most sequenced Gram-positive bacteria, bear the signature indicative of runaway transcription (≥30% identified terminators with stop-to-stem ≤12 nt, Extended Data Fig. 10d-e). Coincidentally, most Firmicutes are resistant to the Rho-inhibitor bicyclomycin[38] and many lack *rho* altogether[39], signifying Rho's diminished role in this phyla (Fig. 4a). We note that the stop-to-stem distance analysis has limited power for some species within Firmicutes that have few identifiable intrinsic terminators, such as *Mycoplasma pneumoniae* in which coupling has been reported[11] and no short-distance terminators were found (Extended Data Fig. 4b-c). Other distant clades of Gram-negative bacteria (e.g., Campylobacterota, Thermotogota) also contain a substantial fraction of short-distance terminators (Fig. 4a), although transcription termination in these

species has not been well-characterized. By contrast, Actinobacteria, Bacteroidetes, and different phyla of Proteobacteria have most terminators far away from the stop codon (Fig. 4a). The prevalence of stop-terminator overlaps across diverse bacterial species suggests that runaway transcription is a common feature in distinct phyla and not unique to *B. subtilis*.

Conclusion

Our understanding of prokaryotic gene regulation has been guided by pioneering examples such as the *E. coli trp* and *his* operons that shed light on an intimate relationship between transcription and translation[14] (Fig. 4b). However, this study shows that in the extensively characterized bacterium *B. subtilis*, transcription and translation are fundamentally disjointed. A much faster RNAP speed not only helps explain the different co-transcriptional regulatory strategies in Firmicutes (Fig. 4c), but also raises important questions regarding mRNA quality control in these bacteria. For example, understanding how *B. subtilis* and related species tolerate accumulation of aberrant transcripts with premature stop codons could provide insights into proteome homeostasis and evolution of new gene functions. More broadly, our results illustrate how a simple kinetic property of the central dogma can profoundly transform the regulatory genome.

Methods

Strains

B. subtilis strains were generated from 168 (*trpC2*) and *E. coli* strains from MG1655. Linearized plasmids, gDNA, and linear PCR products were transformed into 168 using standard protocols relying on natural competence[40]. Sequences of plasmids and integrations were confirmed by Sanger sequencing. All strains are listed in Supplementary Data 6. All plasmids were generated in *E. coli* DH5α cells (except for ORF extension experiments in *E. coli*) using standard protocols and are listed with additional details in Supplementary Data 6.

List of oligonucleotides

Supplementary Data 7 contains the list of oligos used for strain construction, qRT-PCR primers, and probes for Northern blot.

High-throughput expression datasets used (ribosome profiling, Rend-seq)

Available gene expression (ribosome profiling and Rend-seq) datasets were used in the present work: Li et al[41] (accession GSE53767) for ribosome profiling of *E. coli* in rich defined medium; Lalanne et al[35] (accession GSE95211) for ribosome profiling of *B. subtilis* grown in LB, Rend-seq of *B. subtilis* grown in LB (wildtype, Δ*pnpA*, Δ*rho*), Rend-seq of *E. coli* grown in rich defined medium (wildtype, Δ*pnp*, Δ*rnb*), Rend-seq of *V. natriegens* grown in MOPS complete +3 % NaCl, and Rend-seq of *C. crescentus* grown in PYE; DeLoughery et al[42] (accession GSE108295) for Rend-seq of *S. aureus* grown in TSB.

Cell growth

To measure transcription and translation kinetics, pre-cultures were started from single colonies picked into 5 mL LB and grown at 37°C. After 3 h, cultures were diluted to OD_{600}

= 0.005 in 200 mL pre-warmed LB in a 1 L beaker. For alpha-complementation assays, xylose was also added to 200 mL LB to 2% w/v. Cultures were grown with vigorous shaking (>200 rpm) at 37°C. The experiment was performed at OD_{600} = 0.2. For measurements of transcription and translation kinetics in slow growth, cells were grown as above in MOPS Minimal media + 0.4% maltose (100 mL 10X MOPS Mixture, 10 mL 0.132M K_2HPO_4, 20 mL 10% Glutamate, 10 mL 10 mg/mL Tryptophan, 80 mL 5% maltose, 780 mL water).

For Northern blot analysis of *B. subtilis* samples, cells were grown in LB at 37°C with vigorous shaking. *E. coli* samples were grown with additional 50 μg/mL kanamycin and 200 nM anhydrotetracycline to maintain and induce expression from the pSC101-derived plasmid. Pre-cultures were started in pre-warmed LB from fresh colonies. Once OD_{600} reached 0.1, cultures were diluted in the fresh, pre-warmed medium to $OD_{600}=2\times10^{-4}$, and harvested when OD_{600}= 0.3.

For measurements of Rho termination (Fig. 3a, d), single colonies were picked into 5 mL LB and grown for 3 h at 37°C with shaking. Cultures were back diluted OD_{600} = 0.005 into prewarmed 20 mL LB and 1 mM IPTG and grown at 37°C with vigorous shaking. The experiments were performed at OD_{600} = 0.2.

Measurements of protein induction kinetics

Protein induction measurements were based on[8,25,43] with modifications and are detailed below. Cells were grown as described above until they reached an OD = 0.20 – 0.30. Three 1 mL pre-induction cultures were collected into ice-cold stop solution and mixed vigorously before placing on ice. Expression was then induced by mixing the remaining culture with 9 mL IPTG in pre-warmed, shaking LB, such that the final IPTG concentration is 5 mM. Following a brief delay (depending on the length of the construct), 1 mL of culture was collected every 5-10 s for a total of 12 time points. Cells were collected into ice-cold stop solution and mixed vigorously before placing on ice. The stop solution was 50 μL 20 mg/mL chloramphenicol for *E. coli* (based on [25]) and 100 μL buffer containing 10 mg/mL chloramphenicol and 10 mg/mL erythromycin for *B. subtilis*.

For measurements using full length *lacZ*, collected cells were pelleted at 4°C at 18213 g for 4 min and flash frozen in liquid nitrogen. Cell pellets were stored at −80°C until ready to use. For *B. subtilis* measurements, cells were resuspended in 1 mL Z-Buffer containing chloramphenicol, permeabilized with 15 μL toluene and vortexed. 50 μL cells were then added to 450 μL Z-Buffer. For *E. coli* measurements, cells were resuspended in 1.3 mL Z-Buffer containing chloramphenicol and 200 μL cells were then added to 300 μL Z-Buffer. For measurements using *lacZα* fusions, collected cultures were incubated at 37°C for 1 h to allow complementation between LacZω and LacZ*α*. Cells were pelleted at 4°C at 18213 g for 4 min and resuspended in 500 μL Z-Buffer and chloramphenicol, permeabilized with 7.5 μL toluene and vortexed.

Beta-galactosidase activity was measured using a sensitive fluorescent substrate, 4-methylumbelliferyl-D-galactopyranoside (MUG). For all assays, following resuspension in Z-Buffer, 50 μL 2 mg/mL MUG was added, and samples were incubated at 37°C for 30 min before the reaction was stopped with 250 μL of 1 M Na_2CO_3. 50 μL of the total reaction was

moved to a 96-well plate and fluorescence measured on a BioTek Synergy H1 microplate reader with a Blue filter set (EX 360/40 | EM 460/40 | DM 400) with 365 nm excitation and 450 nm emission. To confirm the signal measured was within the linear range of the instrument, induced culture of bGJ74 (full-length *lacZ*) at steady-state were collected and fluorescence of dilutions measured (Extended Data Figure 2a). OD normalized signal measured from the lowest dilution was similar to signal from later time points in induction curves.

Average pre-induction background was subtracted from each point and the square root of signal plotted against time (Schleif plot) to obtain a linear fit with the intercept representing τ_{TL}. Points below background or before signal began to increase consistently were excluded from the fit. For *lacZα*-complementation assays, points after which signal stopped increasing quadratically were excluded and determined as follows. A slope was calculated from the first two points included in the fit and the angle between this line and the next point calculated. This calculation was done for all subsequent pairs of points until an angle >20° between the resulting fit line and the following point was measured. This point and all subsequent points were excluded from the final fit. Error bars reported are the SEM between biological replicates.

In addition to the use of chloramphenicol and erythromycin to stop translation in *B. subtilis*, three additional stop solutions were tested by measuring τ_{TL} for *pycA-lacZα* fusions. Additional stop solutions are as follows. (1) Immediately after collection into chloramphenicol and erythromycin, cells were flash frozen in liquid nitrogen and subsequently thawed in a water bath before complementation as described above. (2) 15 μL toluene was added to chloramphenicol and erythromycin and cells were vortexed immediately after collection. Cells were pelleted and respuspended in 500 μL Z-Buffer and complemented 1 h at 37°C. Cells were then permeablized and assay preformed as described above. Reactions were run for 2.5 h rather than 30 min to account for lower beta-galactosidase activity in these samples. (3) 50 μL 12.5 mg/mL lincomycin was added to chloramphenicol and erythromycin solution. Following collection, assays were performed as described above. The measured τ_{TL} for each of these stop solutions was less than τ_{TL} measured for the stop solution containing only chloramphenicol and erythromycin (Extended Data Figure 2b), suggesting that this stop solution was sufficient to quickly stop translation.

Measurements of mRNA induction kinetics

Cells were grown as described above until they reached an OD = 0.20 – 0.30. A 1 mL pre-induction culture was collected into 1 mL of ice-cold stop solution (60% Ethanol, 2% Phenol pH 8, 10 mM EDTA) and mixed vigorously before placing on ice. Expression was induced as described for protein-induction measurements. Following a brief delay (depending on the length of the construct), 1 mL of cells were collected into 1 mL pre-chilled stop solution (above) on ice every 5-10 s and mixed, for a total of 7 total time points. Cells were pelleted at 4°C at 18213 g for 4 min and flash frozen in liquid nitrogen. Cell pellets were stored at −80°C until ready to use. Cells were lysed in 100 μL TE pH 8 (10 mM Tris-HCl, 1 mM EDTA) containing 10 mg/mL lysozyme for 5 min at 37°C. RNA was then

extracted and gDNA depleted using RNeasy Plus mini kit (Qiagen) following manufacturer's instructions. 1 μg RNA was reverse transcribed with M-MuLV RT and random hexamer priming. RNA was hydrolyzed by adding 10 μL 1M NaOH to the 20 μL RT reaction. The reaction was subsequently neutralized with 10 μL 1M HCl and brought to a final volume of 200 μL with water. qPCR was preformed using 10 μL KAPA SYBR qPCR Master Mix, 6 μL 1 mM primer mix, and 4 μL cDNA and run on a Light Cycler 480 II Real-Time PCR Machine. For full-length *lacZ*, primers annealing to the 3'-end of the transcript were used (Fp: P3105-F, Rp: P3105-R[8]; Supplementary Data 7); for *lacZα* fusion strains, primers at the end of *lacZα* were used (Fp: lacZa-fp, Rp: lacZa-rp; Supplementary Data 7). Resulting Ct values were normalized to time zero and all points with >1.5-fold increases in expression were fit to a line. τ_{TX} was calculated from the intersect of this line and y=1. Error bars reported are the SEM between biological replicates.

Steady-state analysis of mRNA by qRT-PCR

Cells were grown as described above until they reached an OD = 0.2 – 0.3. 1 mL culture was collected into 1 mL ice-cold methanol. Cells were pelleted at 4°C at 18213 g for 4 min and flash frozen in liquid nitrogen. Cell pellets were stored at −80°C until ready to use. RNA was then extracted and gDNA depleted using RNeasy Plus mini kit (Qiagen) following manufacturer's instructions. 1 μg RNA was reverse transcribed with M-MuLV RT and random hexamer priming. RNA was hydrolyzed by adding 10 μL 1M NaOH to the 20 μL RT reaction. The reaction was subsequently neutralized with 10 μL 1M HCl and brought to a final volume of 200 μL with water. qPCR was preformed using 10 μL KAPA SYBR qPCR Master Mix, 6 μL 1 mM primer mix, and 4 μL cDNA and run on a Light Cycler 480 II Real-Time PCR Machine. Steady-state levels of RNA was measured using primers in *lacZα* (Fp: lacZa-fp, Rp: lacZa-rp (Supplementary Data 7)). and normalized to expression of *gyrA* (Fp: gyrA BS qPCR Primer F, Rp: gyrA BS qPCR Primer R (Supplementary Data 7)) to obtain a relative mRNA level.

ORF extension candidate gene selection

The following criteria were considered to select a maximally constraining target for extending the open reading frame into a strong terminator in *B. subtilis*.

1. Stop-to-stem distance larger than 25 nt for the upstream open reading frame.
2. High translation efficiency (75th percentile or higher) of upstream gene as determined by combination of ribosome profiling and Rend-seq (see section "Translation efficiency percentile determination").
3. Efficient intrinsic terminators according to the following criteria:
   a. Less than 1/1000 readthrough fraction as measured by Rend-seq in wild-type, Δ*pnpA*, and Δ*rho*[35].
   b. Strong RNA hairpin secondary structure (ΔG < −17 kcal/mol).
   c. Perfect hairpin (all bases paired in the stem, no bulges).
   d. Long U-tract (7 or more consecutive U residues).

The above criteria were selected to find candidate genes for which transcription and translation could be coupled at the end of the open-reading frame, providing a stringent test of lack of interference of translation on intrinsic transcription termination.

The *pupG* terminator met all the above criterion, was well expressed in our condition, presented a simple operon structure (bicistronic with *drm*) without alternative mRNA isoforms, permitting measurement by Northern blot. Importantly, the open reading frame of *pupG* could be extended all the way inside the loop of its terminator by introducing two single nucleotide mutations. These mutations were at least 20 nucleotides upstream from the start of the hairpin stem, suggesting that they would not affect terminator function. The terminated transcript in the ORF-extended configuration still has a stop codon, avoiding confounding issues with non-stop mediated mRNA degradation.

ORF extension constructs in *B. subtilis*

In order to assess influence of translation on transcription termination[12,27,28,44-47] in *B. subtilis*, we constructed three strains to measure transcriptional readthrough at the *pupG* intrinsic terminator in different contexts: T1+ (*pupG* terminator intact, endogenous context), T1- (*pupG* terminator disrupted: two mutations in the stem resulting in shift from ΔG_{T1+} = −17.5 kcal/mol (free energy of folding of endogenous terminator hairpin) to ΔG_{T1-} = −4.6 kcal/mol (free energy of folding of disrupted terminator hairpin), and one mutation in U-tract: UUUUUUU → UUUCUUU), and ORF extension (*pupG* terminator intact, two upstream mutations to extend the ORF into the terminator hairpin loop). These are illustrated in Extended Data Figure 5a.

In these three different contexts at the *pupG* terminator, a strong intrinsic terminator (*B. subtilis*' endogenous *sodA* intrinsic terminator), labelled T2, was appended downstream (start of the *sodA* terminator region positioned 25 nt downstream of the 3' end of the *pupG* terminator, with 25 nt upstream of the start of the *sodA* hairpin included). This second terminator was placed to capture putative readthrough products from the *pupG* terminator, and allow for measurement of the transcription termination readthrough, analogous to [45].

These modifications were inserted in the endogenous genetic context of *pupG* in *B. subtilis* by appending a downstream co-directional resistance cassette (spectinomycin[48]) not interfering with the upstream and downstream operon, and integrated using long-flanking homology. The constructs were generated by PCR with primers encoding mutations, joined by isothermal assembly, PCR amplified by outer primers and transformed directly in *B. subtilis* following standard protocols relying on natural competence[40]. All strains were confirmed by Sanger sequencing.

ORF extension constructs in *E. coli*

To perform similar measurements in *E. coli*, we transferred the three constructs (T1+, T1-, and ORF extended) on low copy number plasmid pSC101 under the control of TetR repressor with a strong ribosome binding site (derived from pSC101-BD[49], a gift from J. Chin's lab). The full region, from the beginning of the *pupG* gene to downstream of the T2 *sodA* terminator, was amplified by PCR and joined by isothermal assembly to the pSC101 backbone (see Supplementary Data 6 for details). The resulting plasmids were transformed

in *E. coli* K12 MG1655 following the protocol of Chung[50], and assembly junctions were confirmed by Sanger sequencing.

Measurement of transcription readthrough

7.5 mL of culture at OD_{600}=0.3 was added to 7.5 mL of ice-cold methanol, mixed by inversion, and spun down at 3000 rcf for 10 min at 4°C. The supernatant was decanted and the cell pellet frozen at −80°C. RNA was extracted using the RNeasy kit (QIAGEN) following the manufacturer's instruction. 10 μg of total RNA was loaded on an 1.2% TBE (tris-borate-EDTA) agarose gel containing 20 mM guanidine thiocyanate, run for 2 h 30 min at 5 V/cm at 4°C, and transferred to a positively charged nylon membrane by downward capillary transfer. Membranes were cross-linked by UV light, hybridized with short single-stranded DNA probes labeled (T4 PNK, New England Biolabs) with ATP γ-^{32}P (Perkin Elmer), and washed following the manufacturer's instructions. Labelled membranes were exposed to a phosphor storage screen (GE Life Science) for 17 h and imaged with a laser scanner (Typhoon FLA9500, GE Life Science). Bands intensities (background subtracted) were quantified with ImageJ. The probe binding to the region between T1 and T2 had some homology to the endogenous *sodA* region, which led to a strong band running at the expected size for the *sodA* transcript, which provided an additional loading control.

To measure readthrough at the *pupG* terminator in the different translational contexts and strains, two Northern blot probes were used: one upstream in the *pupG* gene for loading and expression control, and one between the T1 and T2 terminators to measure captured transcriptional readthrough products. For each RNA sample (corresponding to one strain), two lanes were loaded with the same RNA and blotted in parallel for the two different probes. The raw Northern blot data is shown in Supplementary Figure 1. The Northern blot was repeated (biological replicate) for the *B. subtilis* constructs with similar results.

As independent confirmation of the Northern blot measurements, we quantified readthrough by qRT-PCR on independent (biological replicates) samples for the various constructs from both *B. subtilis* and *E. coli*. cDNA samples were prepared as described above ("Steady-state analysis of mRNA by qRT-PCR"), and primer pairs quantifying the short (oJBL212+oJBL213 and oJBL209+oJBL210, Supplementary Data 7) and long (oJBL209+oJBL211 and oJBL224+oJBL225, Supplementary Data 7) mRNA isoform from *pupG* terminator were used for readthrough quantification. Readthrough (downstream/upstream signal) was normalized to the T1- construct, leading in *B. subtilis* to T1+/T1- = 0.0006±0.0001, ORF ext./T1- = 0.0023±0.0009, and in *E. coli* T1+/T1- = 0.06±0.04, ORF ext./T1- = 0.7±0.4, where error bars are SEM. from quantification from the four possible primer pairs combinations.

High-confidence list of intrinsic terminators in *E. coli* and *B. subtilis*

To assess the proximity of intrinsic terminators to open reading frames (stop-to-stem distance), we leveraged our set of high-confidence intrinsic terminators set based on end-enriched RNA-seq (Rend-seq) for wildtype and cells deficient in 3'-to-5' exonuclease or Rho[35], with additional quality control criteria described here.

Briefly, the high-confidence terminators were obtained by first mapping *in vivo* RNA 3' ends as positions with a combination of large Z-scores in both 3' peak height and step size. Putative intrinsic terminators within this set of 3' ends were identified based on sequence features (presence of U-tract, and strong nearby upstream RNA hairpin). 3' ends with upstream terminator-like sequences were further filtered based on presence of corresponding 3'-end mapped peak in 3' to 5' exonuclease deletion strains (Δ*pnpA* in *B. subtilis*, Δ*rnb* & Δ*pnp* in *E. coli*). In addition, putative Rho dependent terminators (based on lack of 3'-mapped peak in a Δ*rho B. subtilis* strain, or overlap with bicyclomycin significant transcripts (BSTs) from[16]) were not retained as intrinsic terminators. To be conservative with Rho terminator removal in *E. coli*, we extended the BSTs with an upstream buffer of 300 nt.

Additional criteria were applied to ensure identified intergenic intrinsic terminators were part of simple 3' UTR of their closest upstream gene. First, any terminator sharing the same upstream gene with more than one other terminator were not included (e.g., tandem terminators). Second, any terminator with an intervening 5' end (peak 5'-mapped z score > 12) between its 3' end and the upstream gene was removed to avoid intra-operon regulatory elements such as riboswitches. Third, any terminator for which the average read density (100 nt travelling window) between the terminator and closest upstream stop codon fell below 10% of the read density of the upstream gene was discarded. Finally, terminators for which the stop-to-stem distance was larger than 150 nt were not retained. Our final set in *E. coli* and *B. subtilis* included 409 and 1228 terminators, respectively (see Supplementary Data 2).

The stop-to-stem distance for this high-confidence set (Fig. 2) was determined based on the RNA structure of the hairpin obtained with a constraint of 6 nt unpaired bases from the mapped 3' end (based current understanding of the molecular mechanism of intrinsic termination[51]). This is slightly different with the folding constraint imposed on the purely computationally identified terminators (see section "Classifier for putative intrinsic terminators"), for which all the consecutive U residues (with the end of the U-tract serving as a proxy for the RNA's 3' end) were required to be non-paired. Importantly, we found no strong difference between the distribution of stem-to-stop distances from these two different methods (96% and 91% within 3 nt in *E. coli* and *B. subtilis* respectively).

Correlation between transcription readthrough and stop-to-stem distance

To assess possible interference between translation and intrinsic termination, we compared the terminator readthrough (defined as the fraction of read densities after and before the terminator) measured by Rend-seq for our list of terminators[35] (Supplementary Data 2) to stop-to-stem distances in *E. coli* and *B. subtilis* (Extended Data Figure 5b-i). Readthrough could not reliably be estimated for 17 and 53 terminators in *E. coli* and *B. subtilis* from our final set respectively (either 0 reads downstream or region for estimating readthrough less than 20 nt in size).

In *E. coli*, terminators close to ORFs (stop-to-stem $d \leq 12$ nt) had significantly more readthrough than others (more readthrough corresponds to a poorer terminator). The 30th percentile (q_{30}) of the readthrough distribution for the two sets of terminators (with $d \leq 12$ nt and $d > 12$ nt) has a fold change of $F_{30} := q_{30}^{d \leq 12} / q_{30}^{d > 12} = 3.3$, and $p < 10^{-5}$ (the p-value corresponds to the fraction of random bootstrap sub-samplings of the distribution with

$q_{30}^{d>12} > q_{30}^{d\leq 12}$, *i.e.,* fraction of sub-samplings in which terminators with d>12 nt perform worse than terminators with d≤12 nt), see Extended Data Fig. 5 c,e,g, and i for an illustration of quantities ($q_{30}^{d>12}$, $q_{30}^{d\leq 12}$, F_{30}). The difference remains highly significant when we controlled for possible differences in terminator quality by performing the analysis on terminators with good U-tracts (free energy of RNA/DNA U-tract duplex ΔG_U larger than −5 kcal/mol): F_{30} = 4.0, and $p < 5\times10^{-4}$. In contrast, *B. subtilis*' ORF-proximal intrinsic terminators (stop-to-stem d≤12 nt) does not have significantly more transcription readthrough than other terminators (F_{30} =0.8, p = 0.97; controlling for U-tract (ΔG_U>−5 kcal/mol): F_{30} =1.0, p = 0.35).

We observed similar trends for terminator readthrough measured in the 3' to 5' exonucleases and Rho deletions strains (*B. subtilis* Δ*rho*: F_{30} =1.1, p=0.46; *B. subtilis* Δ*pnpA*: F_{30} =1.2, p=0.16; *E. coli* Δ*rnb*: F_{30} =3.6, $p < 5\times10^{-4}$; *E. coli* Δ*pnp*: F_{30} =2.6, p=0.007; all values listed are controlled for U-tract quality, ΔG_U>−5 kcal/mol).

Fold-change in *rho* null v. WT

Genome-wide fold-change in mRNA levels in Δ*rho* vs. wildtype in *B. subtilis*[52] (green, Fig. 3b) were determined from Rend-seq data (GSE95211) as follows. For each gene, 40 bp on either end of the genes were excluded and the mean read density within the remainder of the gene body calculated in units of read per kilobase per million reads mapped to non rRNAs and tRNAs (rpkm). Reported fold-changes were calculated as the ratio of the rpkm value in Δ*rho* vs. WT. Only genes with more than 50 mapped reads in both wildtype and Δ*rho* were included to avoid counting noise larger than ≈15% (3345 genes in final comparison set). To account for possible slight differences in read depth normalization, the fold-changes were normalized such that the median fold-change was 1 (factor of 1.06). A similar procedure for biological replicates of wildtype was followed (magenta in Fig. 3b) to provide a measure of the reproducibility of the expression quantification using Rend-seq.

Translation efficiency percentile determination

Translation efficiency is the per mRNA initiation rate of ribosomes. The ribosome profiling (deep sequencing of ribosome protected fragments) read density divided by RNA-seq (or Rend-seq) read density for a given gene provides an estimate of translation efficiency for the mRNA corresponding to that gene. For genome-wide distributions and translation efficiency of intact genes, we used previously determined translation efficiency datasets from references[35,41]. For pseudogene regions (Extended Data Figure 7-8, Supplementary Data 3), translation efficiency was estimated by calculating the ribosome profiling and Rend-seq read density over the region normalized by the total number of million reads not mapping to rRNAs and tRNAs (rpkm) and calculating the ratio. The percentile was then determined by comparing to the genome-wide distribution.

Nested antisense RNAs

To look for long untranslated regions in the transcriptome, we searched for non-contiguous operons[53], or excludons[54,55], which we refer to as nested antisense RNAs (asRNAs) here, in *B. subtilis* and *E. coli*. These correspond to transcripts connecting two co-directional genes that are intervened by one or more genes in the opposite direction.

To search for nested asRNAs, we listed all pairs of codirectional genes interrupted by one to three genes in the opposite direction. For each such codirectional gene pair, the moving average (50 nt window) of the read density from Rend-seq data (from wild-type strain) between the midpoint of the two codirectional genes was computed. The codirectional pair was retained if at no position within this range the average read densitd fell below 0.25 read/nt, corresponding to continuous transcription from one gene to another. 50 codirectional gene pairs (45 and 5 with one and two opposite intervening genes, respectively) met this threshold in *B. subtilis*. Interestingly, few codirectional gene pairs (n=3 and with one such pair intervened by a very short 34 a.a. gene at 1st percentile of gene sizes) met this threshold (for a similar total read depth) in *E. coli*, suggesting that codirectional genes intervened by genes on the opposite strand can only rarely be productively co-transcribed in that species.

To focus on untranslated RNAs with functional requirements, we restricted attention to nested asRNAs for which the mean read density between the two co-directional genes (from 50 nt after the end of the first gene to 50 nt before the start of the second gene, 5% winsorized) was equal or larger than half of the read density (5% winsorized) of either the upstream or downstream gene, leading to 35 and 3 final candidates in *B. subtilis* and *E. coli* respectively. These constitute our operational definition of nested asRNAs. The final list of can be found in Supplementary Data 3, with representative examples in *B. subtilis* examples shown in Extended Data Figure 6. In *B. subtilis*, 29/35 of nested asRNAs had less than a 2-fold change in RNA levels (mean Rend-seq read density quantification) upon Rho deletion. In addition to antisense transcripts surrounded by senses genes on both 5’ and 3’ ends (nested), the final list also includes antisense regions inside 3’ and 5’ UTRs of genes (29/35 nested, 4/35 in 5’ UTR, 2/35 in 3’ UTR). In *E. coli*, one of the three identified nested asRNAs was sensitive to bicyclomycin (>2 fold change in mRNA level[16]).

Expressed pseudogenes in *B. subtilis*

To identify expressed pseudogenes *B. subtilis*, we used as starting point all entities annotated as “pseudogenes” (annotation file: GCF_000009045.1_ASM904v1_genomic.gff), leading to 88 hits.

Given that ORF fragments interrupted by frameshifts are annotated as different entities, we clustered annotated pseudogenes spatially with a distance cutoff of 300 bp (end to start), resulting in two types of regions: pseudogene clusters (containing more than one annotated pseudogene) and isolated pseudogenes. With these definitions, *B. subtilis* had 28 pseudogene clusters (comprising 65 pseudogenes) and 23 isolated pseudogenes. Each type was considered separately.

For pseudogene clusters, the following criteria were used to restrict attention to expressed pseudogenes with interrupted translation. First, pseudogene clusters spanning less than 300 nt (from start of first pseudogene to end of last pseudogene in cluster) were not considered (too short, 4/28). Second, pseudogene clusters in which the first annotated pseudogene had read density of less than 0.25 read/nt (Rend-seq) in both wildtype or Δ*rho*, or less than 0.25 read/nt ribosome footprint density (ribosome profiling), were excluded (not measurably expressed, 16/24). Third, pseudogene clusters in which the ribosome footprint density changed by less than twofold from the first to the last pseudogene region were excluded

(lack of translation interruption across cluster, 2/8). In the end, 6/28 pseudogene clusters satisfied these criteria.

For isolated pseudogenes, 15/23 were too short (less than 300 bp in size), and 8/23 were not measurably expressed (less than 0.25 read/nt in both wildtype and Δ*rho* Rend-seq data). 7/23 isolated pseudogenes satisfied both criteria. Note that ribosome profiling expression cutoff was not applied *a priori* for isolated pseudogenes to prevent excluding pseudogenes for which translation was interrupted upstream (see examples below). These were further investigated manually. The following isolated pseudogenes were not retained for polarity assessment: pseudogene *trpC* was not retained as it constituted a full ORF (the allele is rendered non-functional by an in frame 3 bp deletion which disrupts enzyme activity without interrupting translation[56]), *yvzB* constitutes a complete ORF paralogous to flagellar protein *hag* with no evidence of translational disruption, *yvzE* constitutes a paralog of *gtaB* with no evidence of translation disruption (further the pairs contain many identical regions complicating the analysis of quantification based on sequencing). Two isolated pseudogenes consisted of the C-terminal fragments ORFs interrupted by large scale insertions: *spsMc* interrupted by SPβ prophage[57], *sigKc* interrupted by the Skin element[58]. Based on ribosome profiling data, *sigKc* had no clear decrease in translation compared to the upstream gene *yqaB* and was overall very lowly expressed. *sigKc* was thus excluded. Finally, isolated pseudogene *yoyA* had evidence of a plausible short upstream unannotated ORF fragment with overlapping ribosome footprint density. It was thus retained for analysis and the two fragments renamed *yoyAn* and *yoyAc* here. The pseudogene cluster comprising *ydzW* genes had two consecutive nonsense mutations throughout uninterrupted ORFs concomitant with sharp decrease in ribosome footprint density in ribosome profiling. That region was therefore split in three (*ydzWn*, *ydzWm*, *ydzWc*).

Across all pseudogene transcripts, the translation efficiency percentile before and after the nonsense mutation was calculated to assess the decrease in translation (Extended Data Fig. 7, Supplementary Data Table 3).

In the end, we identified 8 expressed pseudogenes with interrupted translation, providing additional independent examples to assess the prevalence of nonsense mediated polarity in *B. subtilis*. Each of these corresponds to a transcribed mRNA across which a sudden drop in translation occurs (experimentally confirmed by a decrease in ribosome footprint density).

We used two metrics to assess Rho-mediated polarity for these pseudogenes. First, we estimated the ratio of the read density in the first and last 15% of the regions to assess progressive decrease in mRNA level (schematically depicted in Extended Data Figure 7a). The ratio spanned the range 0.85 to 1.42 (median 1.29) for the 8 pseudogenes, suggesting lack of Rho-dependent polarity over these regions. Second, the fold-change in read density (subtracting upstream read density corresponding to possible readthrough products) over the full region in Δ*rho* versus wildtype was computed. The fold-changes for considered pseudogenes spanned 0.60 to 1.66 (median 0.70), corresponding to 6th to 80th percentiles in the genome-wide mRNA level fold-change distribution in Δ*rho* versus wildtype.

The lack of large 5’ to 3’ decreasing ramp RNA levels or increase in mRNA levels upon Rho deletion suggests the absence of nonsense mediated polarity for the considered pseudogenes in *B. subtilis*. See Fig. 3b, Extended Data Figure 7 for a visual summary and Supplementary Data 3 for final candidates.

Expressed pseudogene in *E. coli*

The analysis for pseudogenes in *E. coli* was similar to that in *B. subtilis*. There were 199 entities annotated as pseudogenes in *E. coli* (Genbank annotation: U00096.2.gff3). Of these, 66 were in 30 clusters (300 nt distance cutoff for clustering), and the remaining 133 were isolated pseudogenes. 2 pseudogene clusters passed thresholds of expression, ribosome footprint density decrease and size (29/30 >300 bp, 3/29 measurably expressed with cutoff of 0.25 reads/nt in Rend-seq and ribosome profiling, 2/3 with least a 2-fold drop in ribosome footprint read density between the first and last region). 23/133 isolated pseudogenes were both longer than 300 bp (84/133) and measurably expressed (34/133 with >0.25 read/nt Rend-seq read density).

Of expressed and long isolated pseudogenes, 12/23 resulting from insertion element rearrangements were excluded due the difficulty in read-mapping and possible rapid genomic changes near these regions. The remaining 11/23 regions were compared to other *E. coli* strains (O157, GCF_000008865.2; IAI39, GCF_000026345.1; O83, GCF_000183345.1; O104, GCF_000299455.1) to provide context for the possible genetic changes in *E. coli* K-12 MG1655. 6/11 had clear mutations leading to measurably disrupted translation as assessed by ribosome profiling (*ykiA*: ≈2 kb N-terminal portion of gene missing just downstream of promoter compared to O157, *efeU*: 1 bp deletion leading to frameshift in N-terminal portion of the ORF, *gapC*: nonsense mutation ≈100 bp in the gene and 1 bp deletion leading to a stop codon at ≈750 bp, *bcsQ*: nonsense mutation after 6 a.a., *ilvG*: 2 bp deletion leading to stop codon about 1 kb inside the gene, *cybC*: 26 nt deletion leading to the ablation of the beginning of the ORF). 2/11 had frameshift leading early stop codon but no measurable decrease in ribosome footprint density across the stop codon (*yabP* and *yifN*). These were excluded from downstream analysis. Finally, some isolated pseudogenes were excluded because of hard to interpret features. *ybcY*, which has a 2 bp deletion in an ATG either leading to an early stop codon or removal of the start codon (depending on the start codon position), was excluded because the ribosome profiling density was not consistent with either scenarios, suggesting a more complicated situation. *ylbG* was excluded despite clear evidence of decreasing 5’ to 3’ ramp at the transcriptional level given that it constituted an uninterrupted ORF. Finally, *yibJ* (plausibly the C-terminal portion of a longer gene, based on comparison with *E. coli* O157, interrupted by the upstream gene *yibA* in K-12 MG1655) was excluded despite clear evidence of decrease in transcription because of a hard to interpret short RNA nested in the pseudogene body. The pseudogene cluster comprising *gapC* segments had three consecutive nonsense mutations throughout uninterrupted ORFs concomitant with sharp decrease in ribosome footprint density in ribosome profiling. That region was therefore split in four (*gapCn*, *gapCm1*, *gapCm2, gapCc*).

Across all pseudogene transcripts, the translation efficiency percentile before and after the nonsense mutation was calculated to assess the decrease in translation (Extended Data Fig. 8, Supplementary Data Table 3). Note that with substantial polarity (decrease in mRNA level across nonsense mutation), decrease in translation efficiency across the nonsense mutation will naturally be lower (as a reflection of the denominator of the mRNA level being lower).

In total, our final list of expressed pseudogenes with measurable interrupted translation in *E. coli* comprised 8 examples. As in *B. subtilis*, we used decrease in mRNA read density across the pseudogene region as evidence of Rho mediated polarity. The fold-change in RNA read density (Rend-seq) in the first and last 15% of each region was estimated (schematically depicted in Extended Data Figure 8a). 6/8 regions had a fold-change start/end larger than 2 (range: 1.0 to 7.4, median 4.7). In addition, 5/8 showed a fold-change in mRNA levels (either pseudogenes or downstream genes in the same operon) of 2 or more upon treatment with bicyclomycin (an inhibitor of transcription termination factor Rho) compared to untreated control in the study of Peters et al (from the RNA-seq data in Table S2 in reference[16]).

The decreasing mRNA level in a 5' to 3' fashion across the majority of considered pseudogenes with interrupted translation, together with responsiveness to bicyclomycin, confirms that nonsense mediated polarity is common in *E. coli*, consistent with extensive prior literature (reviewed in[59]). See Extended Data Figure 8 for a visual summary, and Supplementary Data 3 of final candidates.

Analysis of C-to-G ratio

To categorize CDSs as either Rho-terminated or non-Rho-terminated, we considered CDSs with at least 150 reads in the Δ*rho* dataset[35], at least 15 reads within the first 10% of the gene in the WT dataset[35], and that were at least 100 bp in length. For each gene meeting these criteria, we calculated four values: the number of reads in the first and last 10% of the gene in WT and Δ*rho* ($reads_{start\_WT}$, $reads_{end\_WT}$, $reads_{start\_Rho}$, $reads_{end\_Rho}$). Rho terminated CDSs were defined as those where $reads_{start\_WT}/ reads_{end\_WT} > 4$ (i.e. exhibited an expression decrease along the gene body) and $reads_{start\_Rho}/ reads_{end\_Rho} < 2$ (to filter out expression decreases resulting primarily from processing or intrinsic termination rather than Rho-dependent termination events). Eleven genes met both of these criteria (*kinB, yhfA, albE, comEC, trpE, msmG, msmE, sqhC, csbX, rapA, ywrK*). *msmG (amyC)* contains several prominent 5' and 3' ends within the first 10% of the gene that are enriched in WT that confound classification of this gene as Rho-terminated. This gene is thus excluded from our analysis. Non Rho-terminated genes were defined as genes where $(reads_{start\_WT}/ reads_{end\_WT})/(reads_{start\_Rho}/ reads_{end\_Rho}) < 1.5$, of which there were 2625 genes. For CDSs in both groups, we calculated the C-to-G ratio for all 100 nt windows and compared the distribution of maximum C-to-G ratios. The distribution of maximum C-to-G ratio for Rho-terminated CDSs was significantly higher than for non Rho-terminated CDSs ($p<10^{-5}$, less than one in 10^5 random sub-samplings (n=10) of non Rho-terminated distribution had higher median maximum C-to-G ratio). See Extended Data Figure 9a.

asRNAs terminated by Rho were identified using the same criteria used to identify Rho terminated CDSs, looking instead in genomic regions antisense to CDSs. Of the 168

asRNAs that passed our expression threshold, 92 were classified as Rho-terminated asRNAs and 112 as non Rho-terminated asRNAs. One of the Rho-terminated asRNAs, that antisense to *cypX*, contained a short transcript in the 5' end of the asRNA driven by a promoter not present in Δ*rho*. This asRNA was thus excluded from analysis. For asRNAs in both groups, we calculated the C-to-G ratio for all 100 nt windows and compared the distribution of maximum C-to-G ratios. The distribution of maximum C-to-G ratio for Rho-terminated asRNAs was significantly higher than for non Rho-terminated asRNAs ($p<10^{-3}$, less than one in 10^3 random sub-samplings (n=10) of non Rho-terminated distribution had higher median maximum C to G ratio compared to sub-sampling (n=10) of Rho-terminated distribution). See Extended Data Figure 9b.

For nested asRNAs, the C-to-G ratio in 100 nt moving window was determined both for antisense regions and for the full region between the co-directional genes. We compared the distribution of maximum C-to-G ratios for the antisense regions within nested asRNAs to the distribution for all regions antisense to CDSs genome-wide. Random subsampling of the genome-wide antisense distribution suggested a highly significant decrease in the maximum C-to-G ratios for nested asRNAs ($p<10^{-5}$, less than one in 10^5 random sub-samplings of the genome-wide distribution had lower median maximum C-to-G ratio). Similar analysis was also performed with the maximum C-to-G ratio for the full region between codirectional genes for the nested asRNAs. The control set was all regions between codirectional with one or two intervening genes in the opposite directions, restricted to the same size range as our set of nested asRNAs (164 to 1606 nt) leading to 484 regions. Again, the maximum C-to-G ratio for nested asRNAs was significantly lower than for the control set ($p<10^{-4}$, less than one in 10^4 random sub-samplings of control distribution had lower median maximum C-to-G ratio).

Species considered for intrinsic terminator identification

Prokaryotic reference and representative genomes from the RefSeq database were downloaded using Assembly from NCBI on 03/16/2019 using query terms:

"Bacteria"[Organism] AND ("representative genome"[refseq category] OR "reference genome"[refseq category]) AND (bacteria[filter] AND latest[filter] AND "complete genome"[filter] AND all[filter] NOT anomalous[filter]). This returned 1648 genomes, which were all searched for intrinsic terminators as described below.

Classifier for putative intrinsic terminators

Each genome (all sequence elements in the reference file: chromosome, plasmids, etc.) was analyzed in isolation. Given the computationally intensive process of RNA secondary structure calculation, putative intrinsic terminators were identified by a two-step process: (1) restrict the attention to U rich regions downstream of genes, and (2) fold upstream RNA and store hairpin characteristics. Stringent selection criteria were then applied on resulting RNA structures and U-tract based on species-specific distribution in properties of RNA secondary structure from randomly selected genome positions.

Specifically, regions of the genome downstream of stop codons (on both strands, with strand-specific information retained), from x_n-10 to x_n+200 nt (where x_n is the annotated

position of the stop codon for ORF n, and + refers to downstream of the gene in the 5' to 3' direction), were retained for all stop codons. If a downstream co-directional gene (n+1) was closer than 200 nt to the stop codons, region x_n-10 to x_{n+1}+30 nt was retained. From within this set of sequences, stretches of more than 5 consecutive T's were identified as putative U-tract of intrinsic terminators for further analysis of upstream secondary structure motif (strong hairpin). For species with GC content exceeding 60%, stretches of 4 consecutive T's were also retained.

For each putative U-tract identified, the minimum free energy RNA secondary structure for the upstream sequences of various lengths (30, 35, 40, 45, 50 nt) ending at the end of the U-tract were obtained using RNAfold (option -C -p)[60], with the constraint that the U residues in the putative U-tract (i.e., only the last stretch of consecutive U residues) remained unfolded. For each structure (and each folded length for a given position) at each putative U-tract, structure properties were extracted: N number of hairpins, S size of stem, L size of loop, ΔG minimum free energy of folding, *l* distance between the 5' end of the putative U-tract and the 3' most base in the stem of the hairpin, and *f* fraction of bases paired in stem.

In order to mitigate differences in GC content across species with regards to selection threshold, and to account for other species-specific differences, we folded 10^4 randomly chosen regions of 40 nt in the genome of each considered species. The hairpin properties for these random regions were stored, as for the U-tract selected regions above.

The selection criteria to identify putative intrinsic terminators were as follows. First, thresholds were applied to geometric features of the folded RNA secondary structure to select for appropriate hairpins: N = 1 (single hairpin for folded region), 5 bp ≤ N_{bp} ≤ 15 bp (sufficient stem), 3 nt ≤ Loop ≤ 8 nt (non-anomalous loop), l=0 nt (stem required to be immediately adjacent to U-tract given the importance of these U residues in termination[61,62]). If for a given U-tract position, more than one folded length hairpin passed these cuts, a single hairpin was selected as follows. Each hairpin was ranked based on three properties: $\Delta G_{hairpin}$, $\Delta G_{hairpin}/N_{bp}$, and f. The selected hairpin for a U-tract was chosen as the hairpin with the highest number of best scoring rank for these properties. In the case of a tie, the hairpin with the highest number of first and second ranks was chosen. In the case of a further tie, the hairpin arising from the shortest folded region was retained.

The final, species-specific, thresholds on the hairpin strength ΔG and fraction of bases paired in the stem f were based on the properties of hairpins from randomly selected regions. For all random regions with N=1 hairpin passing the same geometrical criteria as for putative intrinsic terminators (5 nt ≤ N_{bp} ≤ 15 nt, 3 nt ≤ Loop ≤ 8 nt), the free energy of folding ΔG_1 and ΔG_2 for which less than 1% and 1.5% of randomly folded regions' hairpins obeyed respectively ($\Delta G \leq \Delta G_1$ & $f \geq 0.95$), and ($\Delta G \leq \Delta G_2$ & $f \geq 0.9$) were identified. This determined ΔG_1 and ΔG_2 for each species. If ΔG_1 or ΔG_2 was higher than −6.5 kcal/mol, the free energy threshold was set to −6.5 kcal/mol. Hairpins upstream of U-tract with satisfactory geometrical features (see previous paragraph) further satisfying either ($\Delta G \leq \Delta G_1$ & $f \geq 0.95$) or ($\Delta G \leq \Delta G_2$ & $f \geq 0.9$) were considered putative intrinsic terminators. In the case of species with GC content higher than 60%, the additional threshold of requiring $\Delta G_{hairpin}$ ≤ −20 kcal/mol for hairpins upstream of 4 U residues U-tract was implemented to decrease

the number of false positives (as assessed by our false discovery rate analysis on *C. crescentus*). The procedure is illustrated in Extended Data Figure 10a-c for the case of *V. cholerae*.

Some terminators were further excluded for reasons other than their intrinsic properties to ensure high quality set for stop-to-stem distance assessment. First, terminators arising from genomic elements with size less than 5% of the maximal chromosome size (e.g., small plasmid) were discarded. In addition, instances where multiple terminators were identified downstream of a gene were excluded. Repeated sequence/terminators (terminators with identical sequence and stop-to-stem distances) were excluded. Finally, terminators with stem-to-stop distance larger than 150 nt were excluded to avoid possible annotation errors.

In the end, we identified 301817 terminators for which the stem-to-stop distance could be determined, with a median of 125 per species, and 1434 species with at least 20 identified terminators. A summary for each starting species is in Supplementary Data 4, and all identified terminators satisfying the above criteria are listed in Supplementary Data 5 (we recommend parsing this data file computationally). The Matlab scripts developed for the analysis have been deposited to GitHub (https://github.com/jblalanne/intrinsic_trx_terminator_identifier).

False discovery rate estimation for putative intrinsic terminators

We took advantage of our Rend-seq datasets, which allow for the identification of 3' ends of transcripts, to directly assess the performance of our terminator identification algorithm in *E. coli*, *B. subtilis*, *S. aureus*, *C. crescentus*, and *V. natriegens* (the last species was not included in the RefSeq set, but the algorithm was run on this species as well).

For each species for which we had Rend-seq data, we computed the 3' peak z score[35] in a ± 5 nt neighborhood from the end of the U-tract of putative intrinsic terminators identified by our algorithm. The position was considered to include a 3' end if the maximum peak z score in that neighborhood was above 7. Positions corresponding to insufficiently expressed genes (average read density < 0.25 read/nt in 100 nt window surrounding the position) were excluded from our analysis. For putative intrinsic terminators not containing a 3' end at the end of their U-tract as determined automatically by our peak finding algorithm, we manually verified the presence or absence of clear ends of transcription units. This was important to not confound false positives from the Rend-seq peak finding script with bona fide false positive in the terminator identification algorithm, and was particularly important for the *S. aureus* (library prepared with lower end enrichment) and *C. crescentus* (more noise in read coverage) samples. A position was deemed false positive if the identified position of the terminator was incorrect, or if there was no apparent termination based on our data.

For *C. crescentus* (GC content = 63%), we found that terminators identified using the usual thresholds and 4 U residues U-tract were more frequent false positives. Increasing the stringency of the free energy threshold to $\Delta G \leq -20$ kcal/mol for these 4 U residues U-tract derived hairpins reduced the false positive to the same rate of other putative hairpins. We applied this criterion to all other species with GC content >60%.

We found false positive rates of: 1.1% in *C. crescentus* (2/187), 0.5% in *E. coli* (1/207), 0.1% in *B. subtilis* (1/733), 1.1% in *S. aureus* (2/175), and 0.7% in *V. natriegens* (4/606). Overall, these results suggest an experimentally validated false discovery rate for our terminator identification pipeline of close to or lower than 1% for these five diverse species.

Generation of phylogenetic tree (Fig. 4)

To generate a phylogenetic tree for display of the stop-to-stem distributions for the considered species, we first identified nearly universal and single-copy orthologous genes by using BUSCO (version 3)[63] (lineage file: bacteria_odb9) on each of our RefSeq species in protein assessment mode (-m prot). We only kept for downstream analysis hits labeled as complete, and removed orthologs identified as complete in less than 95% of species, leading to a set of 84 unique orthologs in our final alignment. For each ortholog, the resulting sequences from the RefSeq species were aligned using MAFFT (version 7) (options: --retree 2 –maxiterate 0)[64]. The results for all 84 orthologs were concatenated horizontally. The resulting alignment was trimmed (removing columns with more than 50% gaps or whose consensus frequency was less than 25%). The final multiple sequence alignment contained 20814 positions. In order to avoid double counting identical species, species from the RefSeq whose above alignment was more than 99% identical were grouped together. Specifically, we grouped all members of connected clusters, where species were connected if their above described orthologs' alignment was more than 99% identical. We found 8 connected clusters of more than one species, including a total of 14 species (clusters are identified in Supplementary Data 4 with "grouped w/" flag). These similar species were grouped together for stop-to-stem analysis (Fig. 4 and summary statistic stratification by phylum), after confirming that stop-to-stem distributions were very similar for all grouped species. A single representative species was retained from the sequence alignment for phylogenetic tree generation.

The final phylogenetic tree was generated for all species (with dereplication described above) with at least 20 identified terminators (n=1434) using FastTree (version 2.1.11)[65] with default options on the trimmed concatenated BUSCO MAFFT alignments. The tree was re-rooted at species GCF_000739455.1 and ladderized using phytools[66] (version 0.6-99). The resulting phyla were concordant with those of GTBD89 (only exception: GCF_000217795.1 and GCF_000734015.1 falling between Acquificota and Deferribacterota instead of with other Deltaproteobacteria). Plotting of the tree with stop-to-stem distributions was done in Matlab using custom scripts.

The following GTDB89 phyla[67] were grouped for display in Fig. 4 and phylum-stratified analysis (Extended Data Figure 10d and e: Firmicutes (Firmicutes, Firmicutes_A, Firmicutes_B, Firmicutes_C, Firmicutes_D, Firmicutes_E, Firmicutes_F, Firmicutes_G, Firmicutes_I, Firmicutes_K; the clade within the firmicutes including *Mycoplasma* is shown in Extended Data Figure 4b.), Deltaproteobacteria (Myxococcota, Desulfuromonadota, Desulfobacterota, Desulfobacterota_A), Alphaproteobacteria (Alphaproteobacteria, Magnetococcia), Bacteroides/FCB (Bacteroidota, Gemmatimonadota, Calditrichota, Fibrobacterota), Spirochetes/PVC superphylum (Spirochaetota, Verrucomicrobiota, Verrucomicrobiota_A, Planctomycetota). The Beta and Gammaproteobacteria were

separated in the current work even though they are grouped as a single phylum (Betaproteobacteria) in GTDB89. Other labelled phyla were unique in GTDB89. Other phyla containing few members were not labeled (grey in Fig. 4).

Sequence alignments

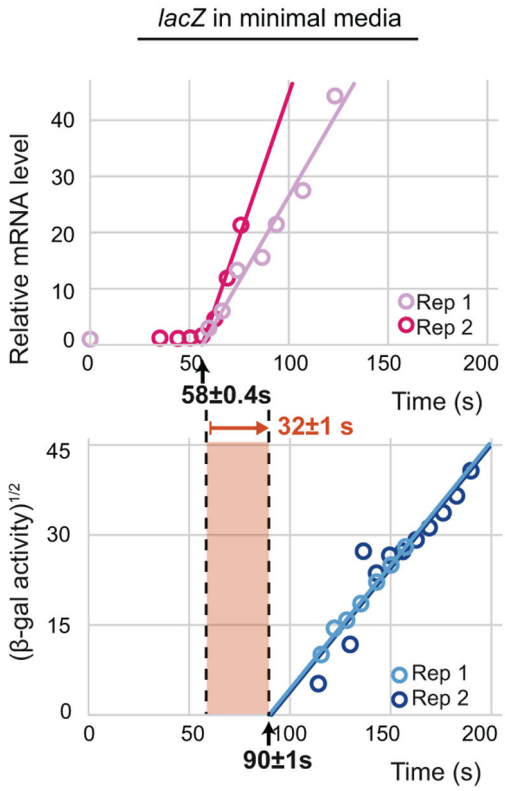
Sequence alignments for NusA, NusG, and RpoB[10] (Extended Data Fig. 4, Supplementary Discussion) were done with MAFFT, trimmed by removing columns at the inflection point of the gap distribution (>95% gaps, such that protein domains present in less than 5% of species considered would not appear in our alignment), and ordered based on the phylogenetic tree of Fig. 4a. For NusA and NusG, domains annotated in Uniprot[68] for the *E. coli* protein were mapped to the positions in the alignment. For RpoB, the conserved sequences identified in[69] (βb1 to βb16) were mapped to positions in the alignment.

Downstream analysis on stop-to-stem distributions

The stop-to-stem distributions in each species, and the grouping by phylum represents two layers of statistical distributions (stop-to-stem distance within a given species, and distribution of the summary statistic of each species' stop-to-stem distribution across species in a given phylum). To highlight the phylum stratification of stop-to-stem distributions of putative terminators, the fraction of species within a phylum for which more than a chosen fraction F of identified terminators had a stem-to-stop distance less than or equal to D was computed for all F and D thresholds (Extended Data Fig. 10d and e). For concreteness, note that Fig. 4 corresponds to thresholds F=30%, and D=12 nt. The fraction of species meeting the threshold on (D, F), as a function of F and D (depicted in Extended Data Figure 10d, thresholds for Fig. 4 F=30%, D=12 nt indicated by yellow star) is then an indication of tolerance of members of the phylum to proximity of intrinsic terminators to coding sequences. We see clear separation in the (D, F) space between different phyla highlighted in Fig. 4., confirming that our conclusion of phylogenetic prevalence of RNAP and ribosome uncoupling are not based on the specific thresholds selected for display in Fig. 4.

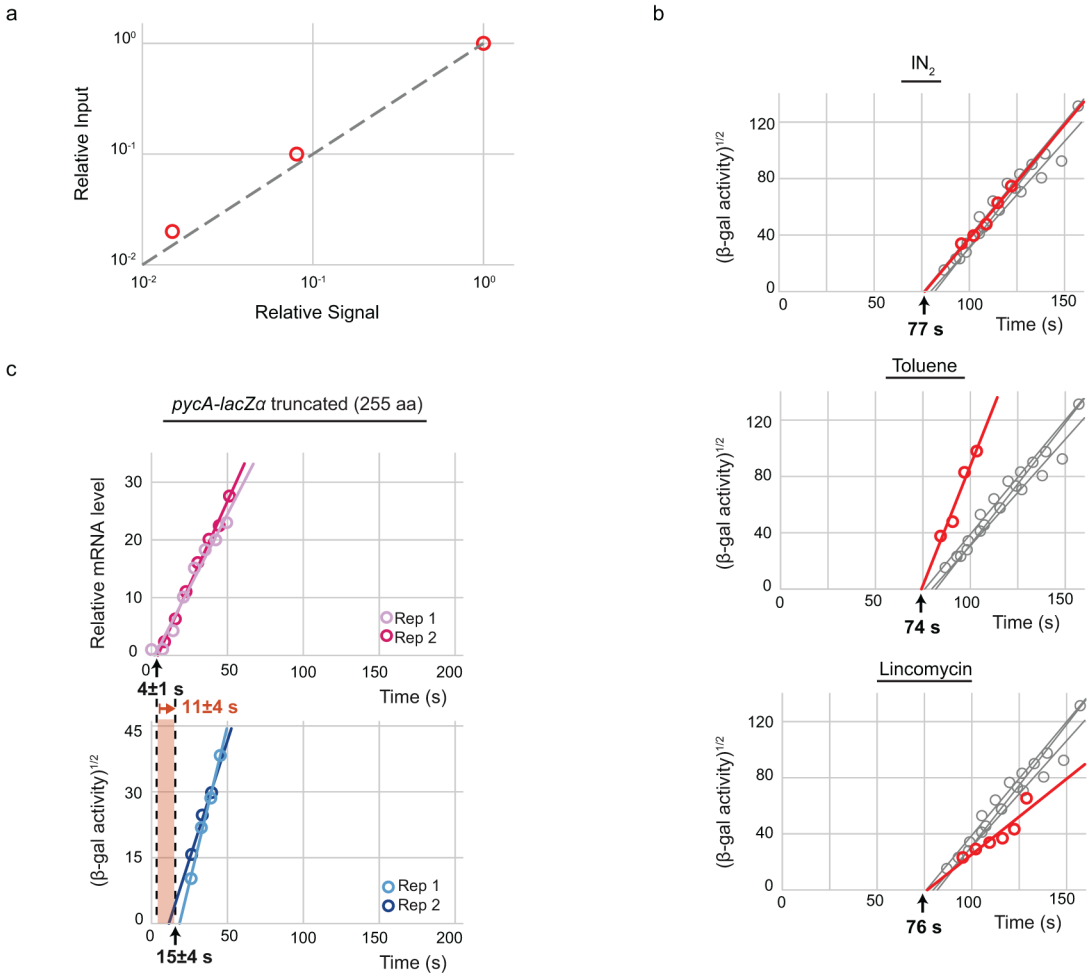
We also confirmed that the differences in stop-to-stem distance distributions were not strongly correlated to other properties, such as overall genome compaction fraction (defined as the fraction of the genome not encoding for genes, which can be both protein coding sequences and non-coding RNAs such as rRNA and tRNAs) and GC content ($R^2 = 0.036$ between median stop-to-stem distance and genome compaction fraction, and $R^2 = 0.122$ between median stop-to-stem distance and GC content).

Extended Data



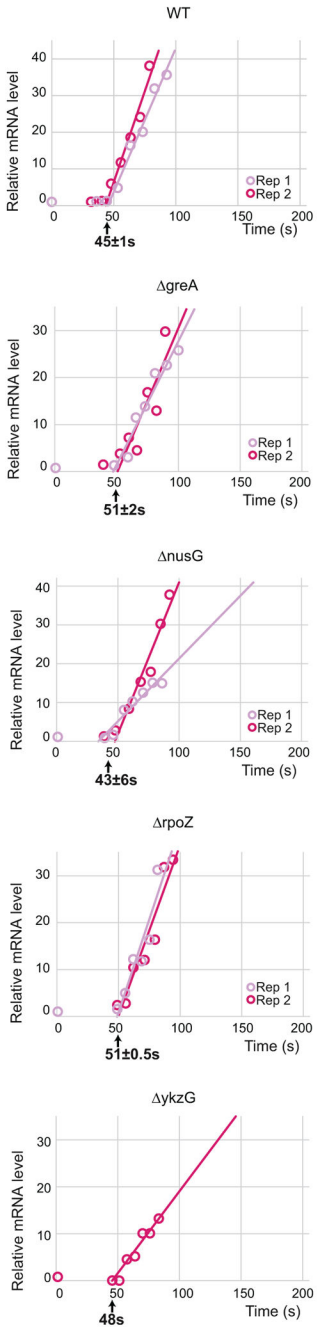
Extended Data Figure 1. Transcription and translation kinetics in slow growth.

Induction time course of *lacZ* mRNA (top) and protein (bottom) as in Fig. 1b, d for WT *B. subtilis* grown in MOPS minimal media + 0.4% maltose (growth rate 0.65 h^{-1}). Lines indicate linear fits after signals rise. Uncertainties are standard error of the mean (SEM) among biological replicates (2).



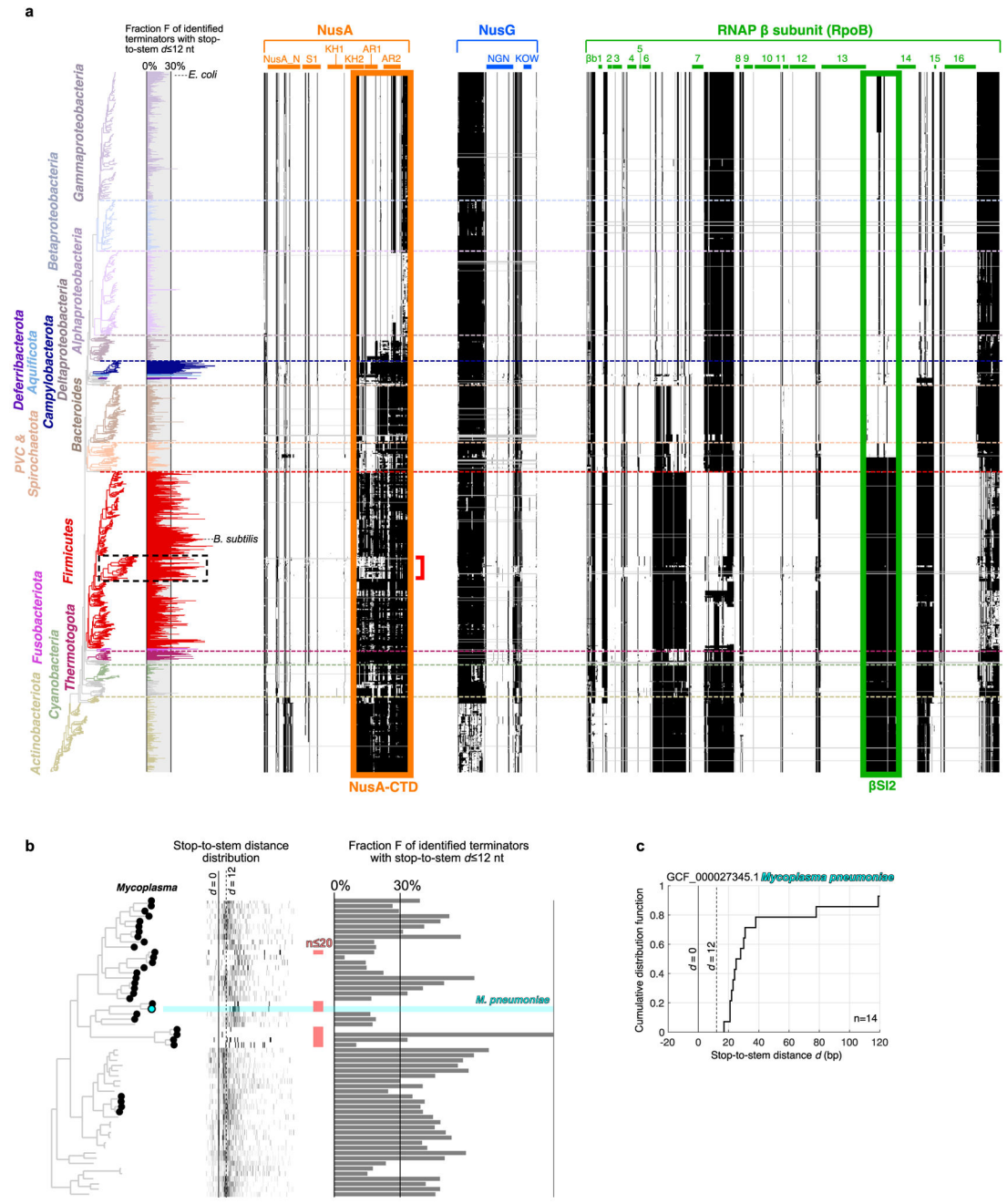
Extended Data Figure 2. Validation of β-gal assay.

a, Measurement of linear range of microplate reader. Fluorescence relative to input of dilutions of an induced culture of bGJ74 (full-length *lacZ*) at steady-state. See Methods. **b,** Effect of different stop solutions on stopping translation. Induction time courses of *pycA-lacZα* protein collected into a stop solution containing chloramphenicol and erythromycin (grey, all plots, from Fig. 1d) or with either flash freezing in liquid nitrogen (top), 15 μL toluene added to the stop solution (middle), or 50 μL 12.5 mg/mL lincomycin added to the stop solution (bottom), shown in red in each plot (as described in Methods). Lines indicate linear fits after signals rise and τ_{TL} is indicated. **c,** Induction time course of truncated *pycA-lacZα* mRNA (top) and protein (bottom) as in Fig. 1b, d. Lines indicate linear fits after signals rise. Uncertainties are standard error of the mean (SEM) among biological replicates (2).



Extended Data Figure 3. Contribution of non-essential RNAP subunits and transcription factors to fast transcription.

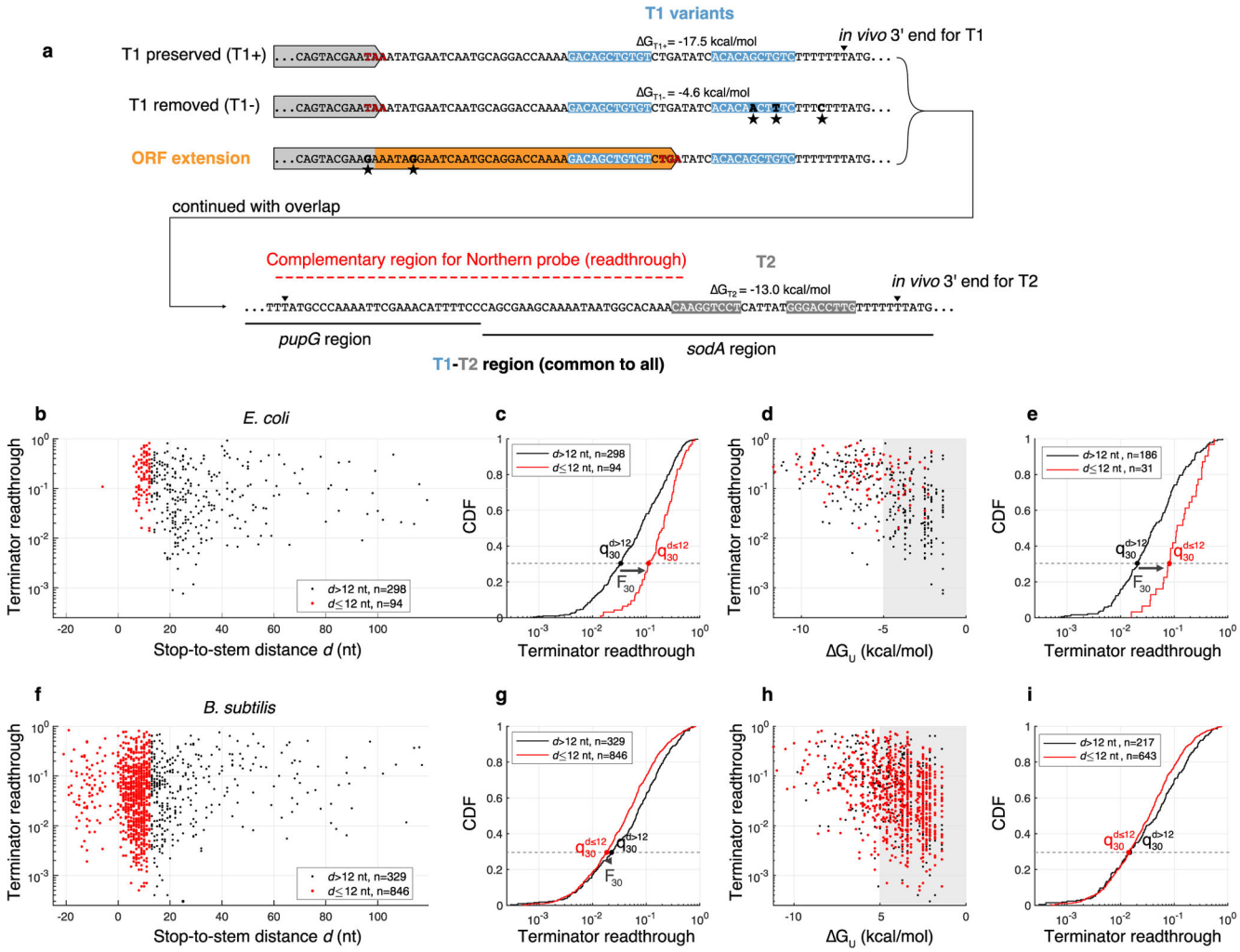
Induction time course of *pycA-lacZα* mRNA in various mutant backgrounds as in Fig. 1b, d. Time course of the same construct in WT from Fig. 1d also shown for reference. Lines indicate linear fits after signals rise. Uncertainties are standard error of the mean (SEM) among biological replicates (1 for Δ*ykzG* and 2 for all others). Time of appearance of full-length mRNA in mutants is not substantially different than that measured in WT (see Supplementary Discussion).



Extended Data Figure 4. Phylogenetic distribution of domain architecture for NusG, NusA and RpoB.

a. Multiple sequence alignments (Methods) for NusA (602 columns), NusG (325 columns), and the β subunit of the RNAP RpoB (1732 columns) for species shown in Fig. 4. The alignments are visualized in a binary fashion to highlight presence/absence of certain domains: white indicates presence of an amino acid in the alignment, and black indicates presence of a gap. The alignments were trimmed by removing columns with >95% gaps. Species with no homologs, partial or pseudogene homologs, or multiple homologs are shown as grey lines. Phylogenetic tree and fraction of terminators with stop-to-stem

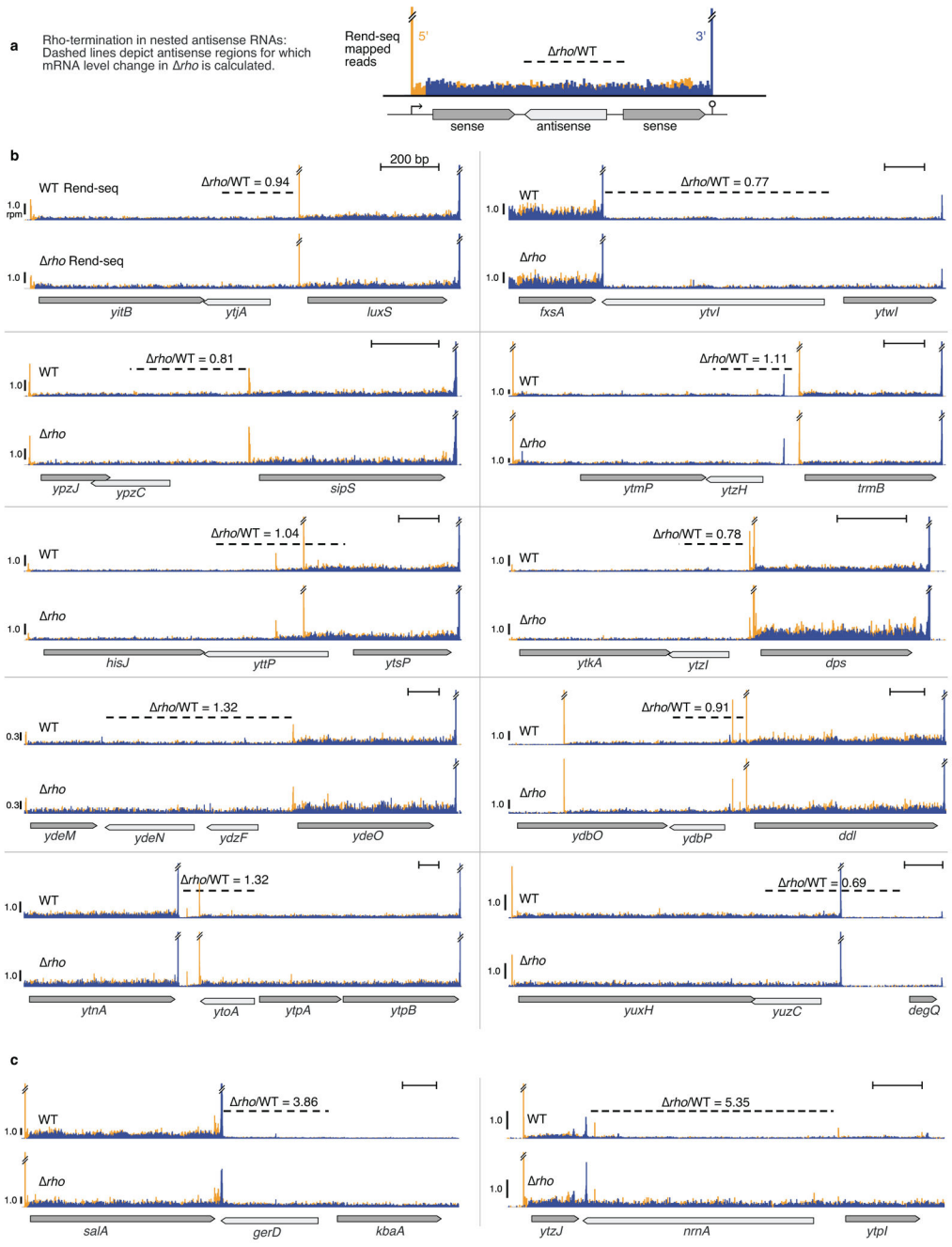
distances within 12 nt from Fig. 4 are reproduced in linearized form. The position of domains from the *E. coli* protein are identified bars above the alignments. For RpoB, conserved bacterial regions identified by[69] (βb1 to βb16) are shown. The NusA C-terminal domain[11,70] (orange box) is missing in a large fraction of Firmicutes (partly present in Mollicutes, which include Mycoplasma and Spiroplasma; red brace), Campylobacterota, Thermotogota, Fusobacteria, and Actinobacteria. NusG has a largely conserved domain architecture, with Actinobacteria showing N-terminal extension. As previously noted in detail[69], the β subunit of the RNAP has multiple insertion domains in diverse bacteria. Insertion domain βSI2, recently implicated[10] (green box) in transcription-translation coupling is lineage-specific and absent in many clades of Gram-positive bacteria, as noted in[10]. Dashed box in tree highlight clade containing *Mycoplasma*. **b.** Close-up view of our analysis of the clade containing *Mycoplasma* (indicated by black dots). Sub-tree includes species with $n \leq 20$ identified terminators (marked in light red). Grayscale representation of stop-to-stem distributions and fraction of terminators with $d \leq 12$ nt are the same as Fig. 4. *M. pneumoniae* is highlighted in cyan, and has no identified terminator (0/14) with $d \leq 12$ nt. **c.** Cumulative distribution of stop-to-stem distance for bioinformatically identified terminators in *M. pneumoniae*.



Extended Data Figure 5. Details of ORF extension constructs and transcription terminator readthrough vs. stop-to-stem distances.

a, Sequence for terminators T1 and T2 for three variants (T1+: *pupG* original terminator, T1-: disrupted *pupG* terminator, ORF extension: original *pupG* with upstream ORF extended inside the loop of the terminator). For T1 and T2, blue and grey shading respectively marks the position of the terminator hairpin stems, with free energy of folding ΔG indicated. Black stars indicate introduced mutations. Downward carets (▼) indicate the position of the 3' ends of associated with intrinsic terminators as determined by Rend-seq. Red dashed line indicates the complementary region of the Northern blot probe to the readthrough product. **b,** Terminator readthrough fraction (defined as the Rend-seq read density after terminator divided by read density upstream of terminator, see [35] for details) as a function of stop-to-stem distance for *E. coli* intrinsic terminators from Fig. 2 for which readthrough could be reliably estimated (n=392). Terminators with stop-to-stem distance $d \leq 12$ nt are highlighted in red. **c,** Cumulative distribution function of terminator readthrough for terminators far (black, $d > 12$ nt) from and close (red, $d \leq 12$ nt) to stop codons. Terminator close to genes have significantly more readthrough (less termination), $p < 10^{-3}$ ($q_{30}^{d>12}$ and $q_{30}^{d\leq 12}$ indicate the 30th percentile in the readthrough distribution for the two categories of terminators, with

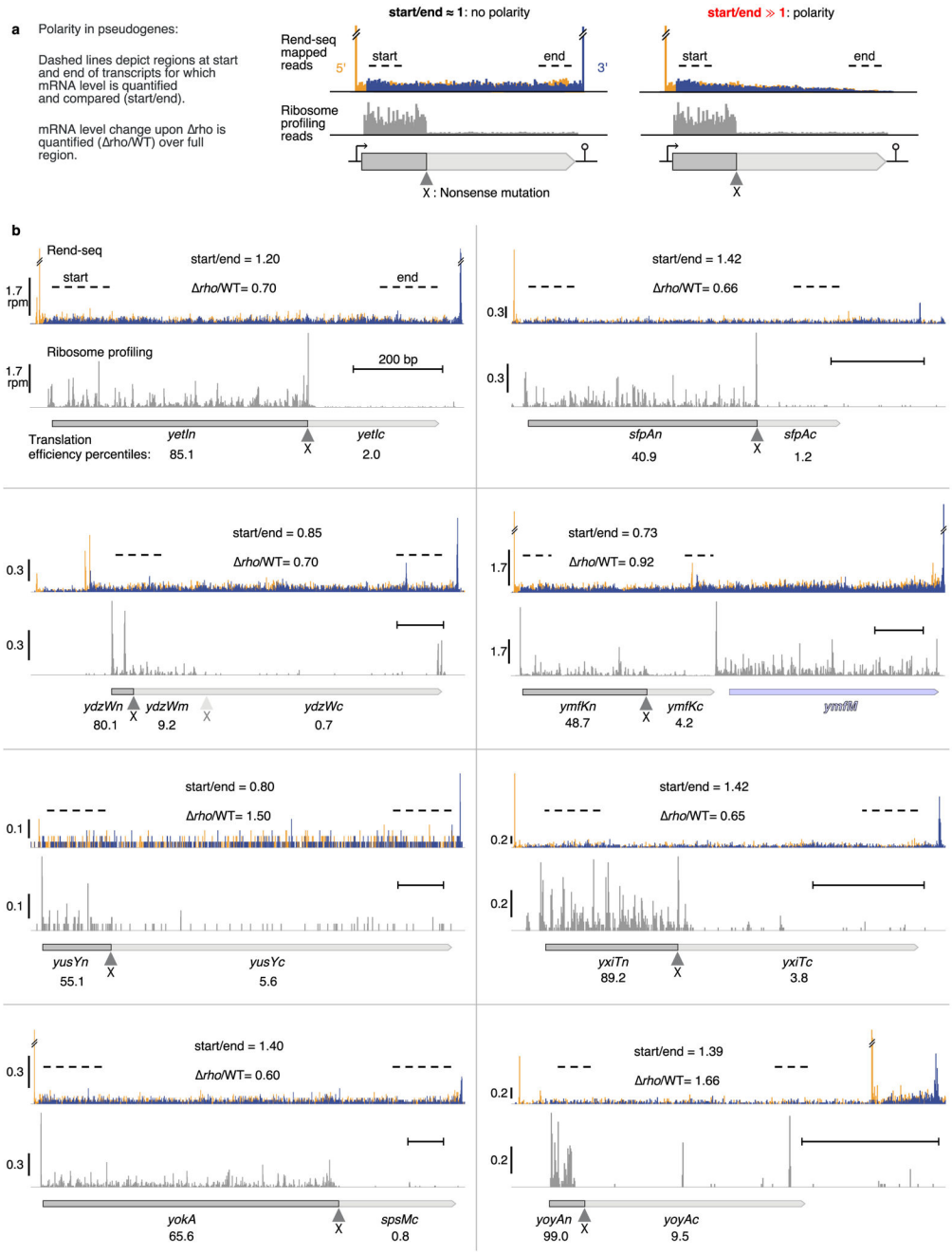
fold-change $F_{30} := q_{30}^{d\leq 12} / q_{30}^{d>12}$, *p*-value determined as the fraction of bootstrap random sub-samplings of the readthrough distributions with $q_{30}^{d>12} > q_{30}^{d\leq 12}$, see Methods) **d,** Terminator readthrough as a function of ΔG_U the U-tract DNA/RNA hybrid free energy (measure of U-tract quality, with larger ΔG_U corresponding to U-rich U-tract). Grey shading indicates cutoff ($\Delta G_U > -5$ kcal/mol) to select good U-tract terminators. **e,** Same as **c**, but restricting to good U-tract terminators, still showing significantly less termination for terminators near ORF, $p<10^{-3}$ (same as above, see Methods). **f-i,** same as **b-e**, but with terminators from *B. subtilis*. Terminators close to ORF do not show less readthrough than their gene-distal counterparts (p>0.3, p-value determined with same strategy as above, see Methods).



Extended Data Figure 6. Examples of identified nested antisense RNAs.

B. subtilis shows a number (n=35, see Methods for selection criteria) of mRNAs with long untranslated regions fully encompassing genes in the antisense directions, which we call nested antisense RNAs (also termed non-contiguous operons[53] or excludons[54]). The majority (n=29/35) of these have a fold-change in mRNA level less than two-fold upon *rho* deletion (Fig. 3b). **a**, Schematic of a nested antisense RNAs with corresponding Rend-seq signal, with orange peaks and blue peaks marking 5' and 3' boundaries of the transcript. **b**, Representative examples of nested antisense RNAs with mRNA level fold change upon *rho* deletion less than 2. Rend-seq data (peak shadows removed, see[35] for details on data

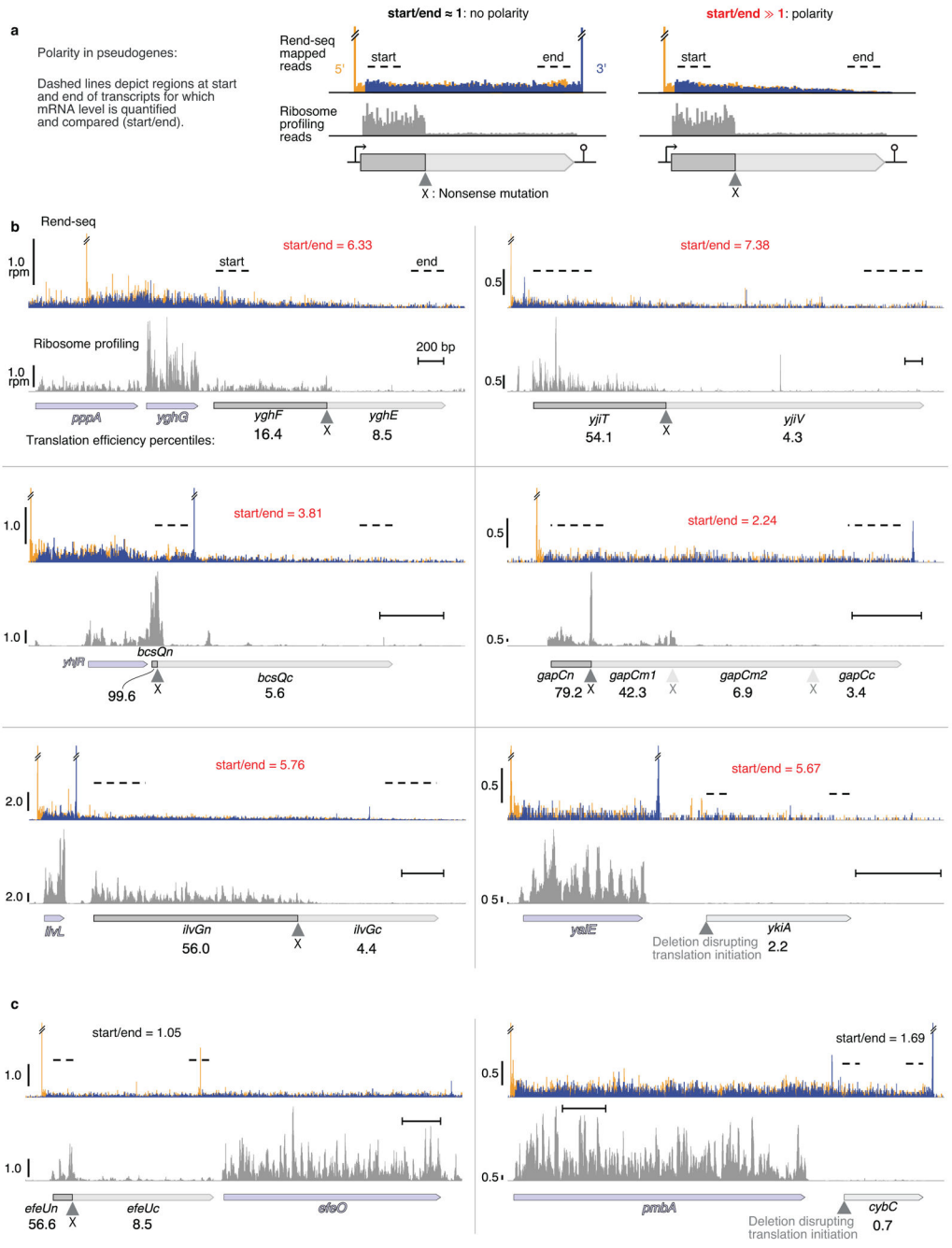
processing) is shown. Orange and blue signal correspond to summed 5'-mapped reads and 3'-mapped reads, respectively (rpm: reads per million). Top trace corresponds to wildtype, and bottom trace to Δ*rho*. Horizontal size marker provides positional scale (200 bp) on each subpanel. Sense and antisense genes are shown in dark and light grey, respectively. Double line breaks (//) indicate truncated Rend-seq signal at peaks. Dashed lines mark regions for which fold-change in read density for Δ*rho*/WT was estimated. The fold-change for each instance is indicated on the graph. **c**, Same as **b**, with representative examples of nested antisense RNAs with increased expression upon *rho* deletion (see Fig. 3b). Three nested antisense RNAs were found in *E. coli* with identical criteria. See Supplementary Data 3 for a list of nested antisense RNAs identified.



Extended Data Figure 7. Expressed pseudogenes with interrupted translation in *B. subtilis* show no polarity.

Expressed pseudogenes endogenously present in the extant genome were used as additional independent experiments to assess the prevalence of Rho-mediated nonsense polarity in *B. subtilis* in situations of obligately uncoupled transcription and translation. Concomitant Rend-seq (mapping operon architecture) and ribosome profiling (measurement of translation) provides stringent data to determine translational status and transcript integrity of mRNAs. **a**, Schematic of analysis: for expressed pseudogenes (see Methods for selection criteria) with translation disruption, polarity was assessed by (1) comparing the mRNA read

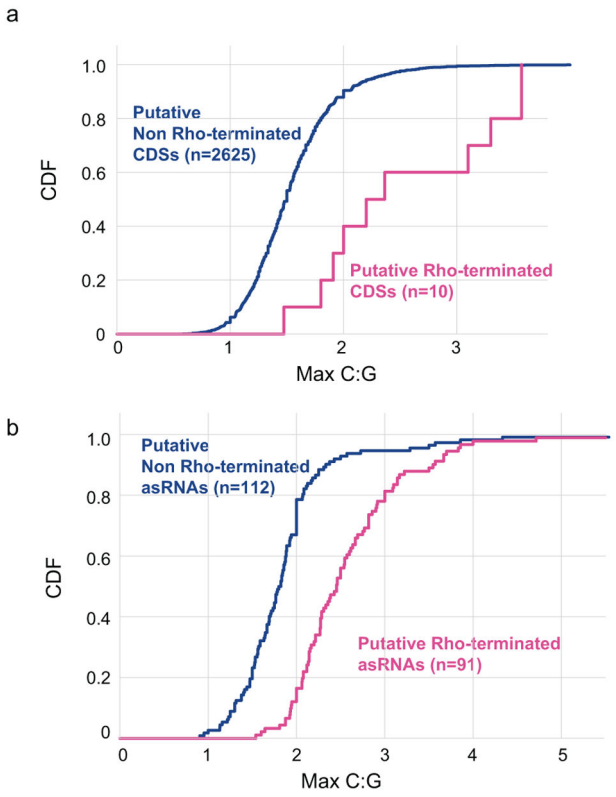
density at start and end of transcription unit, with large changes (start/end≫1) indicative of polarity, and (2) fold change of pseudogene transcript upon *rho* deletion. Position of translation disrupting mutation is shown by ▲ and X. Dark and pale gray indicates region prior and after translation disruption mutation. **b**, Rend-seq and ribosome profiling data for the 8 identified expressed pseudogenes. Each subpanel corresponds to a pseudogene region. Top traces show Rend-seq data (orange and blue signal correspond to summed 5'-mapped reads and 3'-mapped reads, peak shadows removed, see[35] for details on data processing). Orange peaks and blue peaks mark 5' and 3' boundaries of transcripts. Double line breaks (//) indicate truncated Rend-seq signal at peaks. Bottom traces show ribosome profiling data. Translation efficiency (ribosome profiling rpkm/Rend-seq rpkm) percentiles for each pseudogene sub-region (before and after translation disruption) are shown. Horizontal size marker provides positional scale (200 bp) on each subpanel. Nearby intact genes are shown in light blue. rpm: reads per million. Regions used to assess start to end decrease in RNA levels are marked by dashed lines. mRNA levels fold-changes (start/end, and Δ*rho*/WT) are shown. The *ydzW* region showed a second translation disruption the secondary frame, shown as a pale ▲ and X. See Methods, Fig. 3b and Supplementary Data 3 for details.



Extended Data Figure 8. Most expressed pseudogenes with interrupted translation in *E. coli* show polarity.

Similar to Extended Data Figure 7. Expressed pseudogenes endogenously present in the extant genome were used as additional independent experiments to assess the prevalence of Rho-mediated nonsense polarity in *E. coli* in situations of obligately uncoupled transcription and translation. Concomitant Rend-seq (mapping operon architecture) and ribosome profiling (monitoring translation) provides stringent data to determine translational status and transcript integrity on mRNAs. **a**, Schematic of analysis: for expressed pseudogenes (see Methods for selection criteria) with translation disruption, polarity was assessed by

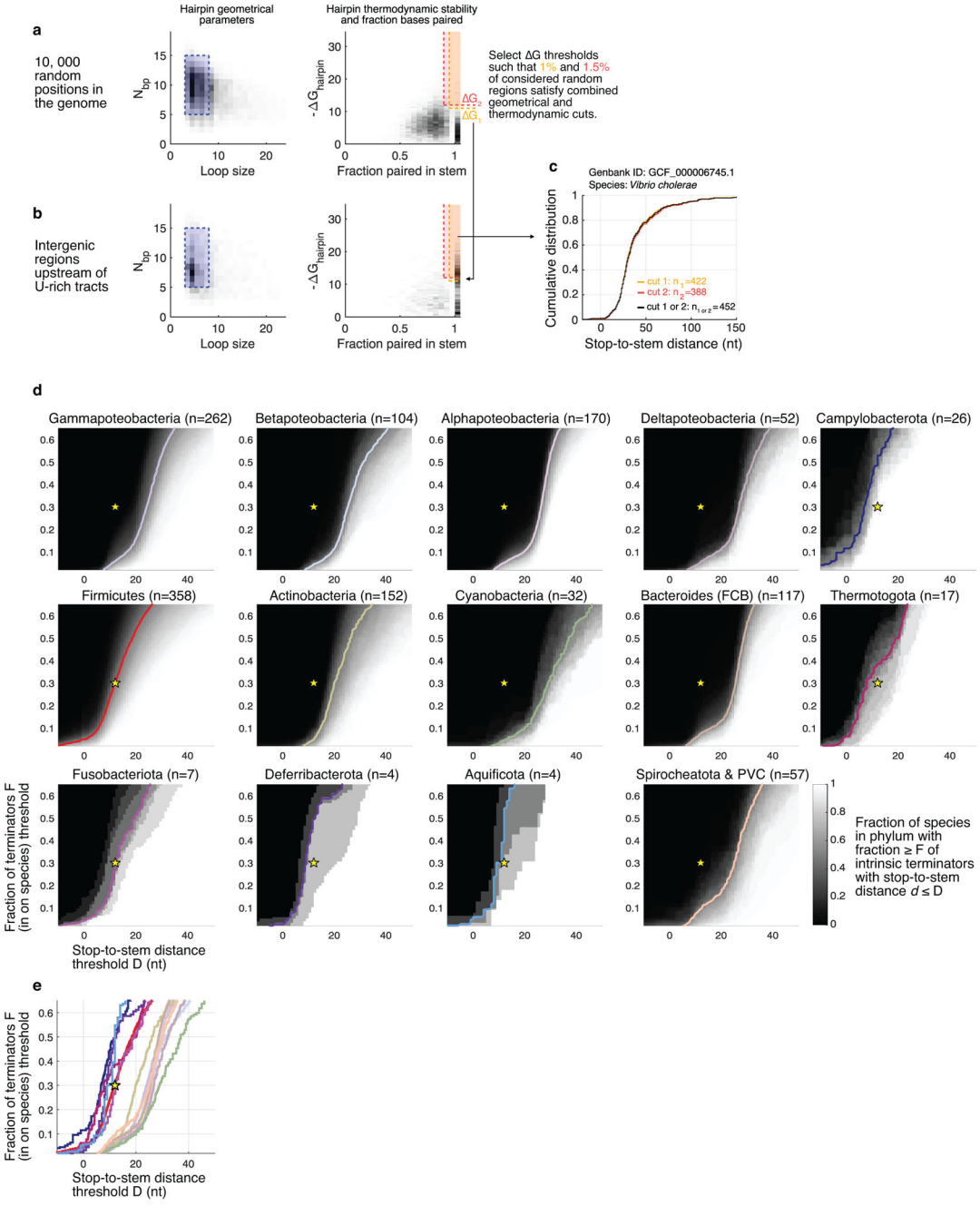
comparing the mRNA read density at start and end of transcription unit, with large changes (start/end≫1) indicative of polarity. **b**, Rend-seq and ribosome profiling data for the identified expressed pseudogene with evidence of polarity. Each subpanel corresponds to a pseudogene region. Top traces correspond to Rend-seq data (orange and blue signal correspond to summed 5'-mapped reads and 3'-mapped reads, peak shadows removed, see[35] for details on data processing). Orange peaks and blue peaks mark 5' and 3' boundaries of transcripts. Double line breaks (//) indicate truncated Rend-seq signal at peaks. Bottom traces show ribosome profiling data. Translation efficiency (ribosome profiling rpkm/Rend-seq rpkm) percentiles for each pseudogene sub-region (before and after translation disruption) are shown. Horizontal size marker provides positional scale (200 bp) on each subpanel. Light blue arrows correspond to nearby intact genes. rpm: reads per million. Regions used to assess start to end decrease in RNA levels are marked by dashed lines. mRNA levels fold-changes (start/end) are shown. The *gapC* region showed sequential translation disruptions secondary frames, shown as a pale ▲ and X. **c**, same as **b**, but for the two cases with no evidence of polarity. The translation disruptions mutation in *ykiA* and *cybC* are deletion of the beginning of ORFs. See Methods, Fig. 3b and Supplementary Data 3 for details.



Extended Data Figure 9. Analysis of C-to-G ratio for putative Rho-terminated RNAs.

a, Cumulative distributions of maximum C-to-G ratio ("Max C:G") of 100 nt sliding windows within non Rho-terminated coding sequences (CDSs, blue, n=2625) and Rho-terminated CDSs (magenta, n=10). Median of Max C:G is higher for Rho-terminated CDSs (magenta) than non Rho-terminated CDSs (blue) ($p<10^{-5}$, less than one in 10^5 random sub-

samplings (n=10) of non Rho-terminated distribution had higher median maximum C-to-G ratio). **b,** Cumulative distributions as in **a** for asRNAs that are not terminated by Rho (blue, n=112) and asRNAs that are terminated by Rho (magenta, n=91). Median of Max C:G is higher for Rho-terminated asRNAs than non Rho-terminated asRNAs ($p<10^{-3}$, less than one in 10^3 random sub-samplings (n=10) of non Rho-terminated distribution had higher median maximum C to G ratio compared to sub-sampling (n=10) of Rho-terminated distribution). See Methods.



Extended Data Figure 10. Illustration of terminator identification pipeline and analysis of stem-to-stop distribution stratified by phyla.

The terminator identification pipeline selects for strong hairpins immediately upstream of long U-tract found downstream of genes. Thresholds on hairpin folding free energy are determined on a species-by-species basis based on properties of randomly selected regions in respective genomes. The case of *V. choloerae* is illustrated in **a**-**c**. **a,** Results of folding 10^4 regions of 40 nt chosen at random positions in the genome. Left panel shows the 2D distribution as a heatmap (dark positions corresponding to more density) of hairpin geometrical parameters (number of base pairs in stem N_{bp}, length of loop). Geometric thresholds are highlighted with blue dashes (5 bp ≤ N_{bp} ≤ 15 bp, 3 nt ≤ Loop ≤ 8 nt) and retained region by blue shading. Right panel shows the 2D distribution as a heatmap (dark positions correspond to more density) of hairpin free energy of folding $\Delta G_{hairpin}$ and fraction of bases paired in stem *f*. Thresholds ΔG_1 and ΔG_2 on $\Delta G_{hairpin}$ are chosen such the total fraction of hairpin from random regions meeting geometrical (blue shading in left panel) and thermodynamic thresholds are 1% (orange, $\Delta G_{hairpin} \leq \Delta G_1$ and $f \geq 0.95$) and 1.5% (red, $\Delta G_{hairpin} \leq \Delta G_2$ and $f \geq 0.9$). **b,** Similar as for **a**, but for regions seeded by U-tracts (stretch of 5 or more consecutive T's in the genome downstream of genes). Note the excess density of hairpins with strong energy of folding and large fraction of bases paired, corresponding to putative intrinsic terminators. **c,** Distribution of stop-to-stem distances for terminators passing thresholds shown in **b**. See Supplementary Data 2, Supplementary Data 3, and Methods for details of computational pipeline. **d** and **e**, Phylum stratified analysis on the stop-to-stem distribution. **d,** Each subpanel shows as a 2D greyscale the fraction of species within each phylum (shown in Fig. 4) for which more than fraction F (y-axis) of terminators have stop-to-stem distances less than or equal to D (x-axis). Black regions correspond to no species in the phylum, white all species. The contour line in the (D,F) space marks points where 50% of species in the phylum have fraction ≥F of their terminators with stop-to-stem distance ≤D. The yellow stars mark the thresholds used in Fig. 4 (D=12 nt, F=30%). For example, about 50% of species analyzed in the Firmicutes have more than 30% of their terminators within 12 nt of upstream ORF (red contour line intersecting yellow star). **e,** The 50% species contour lines from **d** reported to the same panel, showing clear separation between phyla.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

We thank James Taggart, Matthew Tien, and A. Grossman's lab for providing plasmids; Dylan McCormick for help with strain generation, Lydia Herzel, J. Taggart, and M. Tien for help collecting cultures for measurements of induction kinetics; members of the G.-W.L. and A. Grossman Labs for discussions; L. Herzel, D. J. Parker, A. Grossman, J. Peters and V. Siegel for comments on the manuscript; members of the BioMicroCenter at MIT for help in performing the genomic DNA sequencing.

Funding: This research was supported by NIH R35GM124732, Pew Biomedical Scholars Program, a Sloan Research Fellowship, Searle Scholars Program, the Smith Family Award for Excellence in Biomedical Research, an NSF graduate research fellowship (to G.E.J.), an NIH Pre-Doctoral Training Grant (T32 GM007287, to G.E.J. and M.L.P.), an NSERC graduate fellowship (to J.B.L.), and an HHMI International Student Fellowship (to J.B.L.).

Main references:

1. Adhya S & Gottesman M Control of Transcription Termination. Annu. Rev. Biochem 47, 967–996 (1978). [PubMed: 354508]
2. Richardson JP Preventing the synthesis of unused transcripts by rho factor. Cell 64, 1047–1049 (1991). [PubMed: 2004415]
3. Landick R, Carey J & Yanofsky C Translation activates the paused transcription complex and restores transcription of the trp operon leader region. Proc. Natl. Acad. Sci. USA 82, 4663–4667 (1985). [PubMed: 2991886]
4. Proshkin S, Rahmouni AR, Mironov A & Nudler E Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328, 504–508 (2010). [PubMed: 20413502]
5. Burmann BMB et al. A NusE: NusG complex links transcription and translation. Science 328, 501–504 (2010). [PubMed: 20413501]
6. Kohler R, Mooney RA, Mills DJ, Landick R & Cramer P Architecture of a transcribing-translating expressome. Science 356, 194–197 (2017). [PubMed: 28408604]
7. Fan H et al. Transcription–translation coupling: direct interactions of RNA polymerase with ribosomes and ribosomal subunits. Nucleic Acids Res. 45, 11043–11055 (2017). [PubMed: 28977553]
8. Zhu M, Mori M, Hwa T & Dai X Disruption of transcription-translation coordination in Escherichia coli leads to premature transcriptional termination. Nat. Microbiol 4, 2347–2356 (2019). [PubMed: 31451774]
9. Webster MW et al. Structural basis of transcription-translation coupling and collision in bacteria. bioRxiv 1–32 (2020).
10. Wang C et al. Structural basis of transcription-translation coupling. bioRxiv 21, 1–9 (2020).
11. O'Reilly FJ et al. In-cell architecture of an actively transcribing-translating expressome. bioRxiv 1–15 (2020).
12. Roland KL, Liu C & Turnbough CL Role of the ribosome in suppressing transcriptional termination at the pyrBI attenuator of Escherichia coli K-12. Proc Natl Acad Sci USA 85, 7149–7153 (1988). [PubMed: 2459698]
13. Yanofsky C Attenuation in the control of expression of bacterial operons. Nature 289, 751–758 (1981). [PubMed: 7007895]
14. Landick R, Turnbough C & Yanofsky C Transcription attenuation in Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology (eds. Neidhardt F et al.) 1276–1301 (American Society for Microbiology, 1996).
15. Kervestin S & Jacobson A NMD: A multifaceted response to premature translational termination. Nat Rev Mol Cell Biol. 13, 700–712 (2012). [PubMed: 23072888]
16. Peters JM et al. Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes Dev. 26, 2621–2633 (2012). [PubMed: 23207917]
17. Henkin TM Control of transcription termination in prokaryotes. Annu. Rev. Genet 30, 35–57 (1996). [PubMed: 8982448]
18. Winkler WC & Breaker RR Regulation of Bacterial Gene Expression By Riboswitches. Annu. Rev. Microbiol 59, 487–517 (2005). [PubMed: 16153177]
19. Babitzke P & Yanofsky C Reconstitution of Bacillus subtilis trp attenuation in vitro with TRAP, the trp RNA-binding attenuation protein. Proc. Natl. Acad. Sci. USA 90, 133–137 (1993). [PubMed: 7678334]
20. Ingham CJ, Dennis J & Furneaux PA Autogenous regulation of transcription termination factor Rho and the requirement for Nus factors in Bacillus subtilis. Mol. Microbiol 31, 651–663 (1999). [PubMed: 10027981]
21. Hidenori S & Henner DJ Construction of a single-copy integration vector and its use in analysis of regulation of the trp operon of Bacillus subtilis. Gene 43, 85–94 (1986). [PubMed: 3019840]
22. Yakhnin H, Babiarz JE, Yakhnin AV & Babitzke P Expression of the Bacillus subtilis trpEDCFBA operon is influenced by translational coupling and rho termination factor. J. Bacteriol 183, 5918–5926 (2001). [PubMed: 11566991]

23. Schleif R, Hess W, Finkelstein S & Ellis D Induction kinetics of the L arabinose operon of Escherichia coli. J. Bacteriol 115, 9–14 (1973). [PubMed: 4577756]
24. Vogel U & Jensen KF The RNA chain elongation rate in Escherichia coli depends on the growth rate. J. Bacteriol 176, 2807–2813 (1994). [PubMed: 7514589]
25. Dai X et al. Reduction of translating ribosomes enables Escherichia coli to maintain elongation rates during slow growth. Nat. Microbiol 2, 1–9 (2016).
26. Artsimovitch I, Svetlov V, Anthony L, Burgess RR & Landick R RNA polymerases from Bacillus subtilis and Escherichia coli differ in recognition of regulatory signals in vitro. J. Bacteriol 182, 6027–6035 (2000). [PubMed: 11029421]
27. Li R, Zhang Q, Li J & Shi H Effects of cooperation between translating ribosome and RNA polymerase on termination efficiency of the Rho-independent terminator. Nucleic Acids Res. 44, 2554–2563 (2015). [PubMed: 26602687]
28. Wright JJ & Hayward RS Transcriptional termination at a fully rho-independent site in Escherichia coli is prevented by uninterrupted translation of the nascent RNA. EMBO J. 6, 1115–1119 (1987). [PubMed: 3036492]
29. Reilman E, Mars RAT, Van Dijl JM & Denham EL The multidrug ABC transporter BmrC/BmrD of Bacillus subtilis is regulated via a ribosome-mediated transcriptional attenuation mechanism. Nucleic Acids Res. 42, 11393–11407 (2014). [PubMed: 25217586]
30. Dar D et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science 352, aad9822 (2016). [PubMed: 27120414]
31. Yakhnin H et al. Nusg-dependent rna polymerase pausing and tylosin-dependent ribosome stalling are required for tylosin resistance by inducing 23s rrna methylation in bacillus subtilis. MBio 10, 1–14 (2019).
32. Goodson JR, Klupt S, Zhang C, Straight P & Winkler WC LoaP is a broadly conserved antiterminator protein that regulates antibiotic gene clusters in Bacillus amyloliquefaciens. Nat. Microbiol 2, 1–10 (2017).
33. Garza-Sánchez F, Schaub RE, Janssen BD & Hayes CS tmRNA regulates synthesis of the ArfA ribosome rescue factor. Mol. Microbiol 80, 1204–1219 (2011). [PubMed: 21435036]
34. Shimokawa-Chiba N et al. Release factor-dependent ribosome rescue by BrfA in the Gram-positive bacterium Bacillus subtilis. Nat. Commun 10, 5397 (2019). [PubMed: 31776341]
35. Lalanne JB et al. Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. Cell 173, 749–761 (2018). [PubMed: 29606352]
36. Gowrishankar J & Harinarayanan R Why is transcription coupled to translation in bacteria? Molecular Microbiology 54, 598–603 (2004). [PubMed: 15491353]
37. Nicolas P et al. Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. Science 335, 1103–6 (2012). [PubMed: 22383849]
38. Nishida M, Mine Y, Matsubara T, Goto S & Kuwahara S Bicyclomycin, a new antibiotic: In vitro and in vivo antimicrobial activity. J. Antibiot. (Tokyo) 25, 582–593 (1972). [PubMed: 4567453]
39. D'Heygère F, Rabhi M & Boudvillain M Phyletic distribution and conservation of the bacterial transcription termination factor rho. Microbiol. (United Kingdom) 159, 1423–1436 (2013).

Methods References

40. Harwood CR and Cutting SM Molecular Biological methods for Bacillus. Molecular Biological Methods for Bacillus (John Wiley, 1990).
41. Li G-W, Burkhardt D, Gross C & Weissman JS Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell 157, 624–635 (2014). [PubMed: 24766808]
42. DeLoughery A, Lalanne J-B, Losick R & Li G-W Maturation of polycistronic mRNAs by the endoribonuclease RNase Y and its associated Y-complex in Bacillus subtilis. Proc. Natl. Acad. Sci. USA 115, E5585–E5594 (2018). [PubMed: 29794222]
43. Zhu M, Dai X & Wang Y-P Real time determination of bacterial in vivo ribosome translation elongation speed based on LacZα complementation system. Nucleic Acids Res. 44, gkw698 (2016).

44. Bonekamp F, Clemmesen K, Karlstrom O & Jensen KF Mechanism of UTP-modulated attenuation at the pyrE gene of Escherichia coli: an example of operon polarity control through the coupling of translation to transcription. Embo J 3, 2857–2861 (1984). [PubMed: 6098450]
45. Abe H, Abo T & Aiba H Regulation of intrinsic terminator by translation in Escherichia coli: Transcription termination at a distance downstream. Genes to Cells 4, 87–97 (1999). [PubMed: 10320475]
46. Unniraman S, Prakash R & Nagaraja V Alternate Paradigm for Intrinsic Transcription Termination in Eubacteria. J. Biol. Chem 276, 41850–41855 (2001). [PubMed: 11551936]
47. Kobayashi K, Kuwana R & Takamatsu H kinA mRNA is missing a stop codon in the undomesticated Bacillus subtilis strain ATCC 6051. Microbiology 154, 54–63 (2008). [PubMed: 18174125]
48. Guérout-Fleury AM, Shazand K, Frandsen N & Stragier P Antibiotic-resistance cassettes for Bacillus subtilis. Gene 167, 335–336 (1995). [PubMed: 8566804]
49. Rackham O & Chin JW A network of orthogonal ribosome·mrna pairs. Nat. Chem. Biol 1, 159 – 166 (2005). [PubMed: 16408021]
50. Chung CT, Niemela SL & Miller RH One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution. Proc. Natl. Acad. Sci. USA 86, 2172–2175 (1989). [PubMed: 2648393]
51. Peters JM, Vangeloff AD & Landick R Bacterial Transcription Terminators: The RNA 3′-End Chronicles. J. Mol. Biol 412, 793–813 (2011). [PubMed: 21439297]
52. Bidnenko V et al. Termination factor Rho: From the control of pervasive transcription to cell fate determination in Bacillus subtilis. PLoS Genetics 13, (2017).
53. Sáenz-Lahoya S et al. Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. Proc. Natl. Acad. Sci. USA 116, 1733–1738 (2019). [PubMed: 30635413]
54. Sesto N, Wurtzel O, Archambaud C, Sorek R & Cossart P The excludon: A new concept in bacterial antisense RNA-mediated gene regulation. Nat. Rev. Microbiol 11, 75–82 (2013). [PubMed: 23268228]
55. Yan B, Boitano M, Clark TA & Ettwiller L SMRT-Cappable-seq reveals complex operon variants in bacteria. Nat. Commun 9, (2018).
56. Albertini AM & Galizzi A The sequence of the trp operon of Bacillus subtilis 168 (trpC2) revisited. Microbiology 145, 3319–3320 (1999). [PubMed: 10627030]
57. Abe K et al. Developmentally-Regulated Excision of the SPβ Prophage Reconstitutes a Gene Required for Spore Envelope Maturation in Bacillus subtilis. PLoS Genet. 10, e1004636 (2014). [PubMed: 25299644]
58. Stragier P, Kunkel B, Kroos L & Losick R Chromosomal rearrangement generating a composite gene for a developmental transcription factor. Science 243, 507–512 (1989). [PubMed: 2536191]
59. Ciampi MS Rho-dependent terminators and transcription termination. Microbiology 152, 2515–2528 (2006). [PubMed: 16946247]
60. Tafer H et al. ViennaRNA Package 2.0. Algorithms Mol. Biol 6, 26 (2011). [PubMed: 22115189]
61. Gusarov I & Nudler E The mechanism of intrinsic transcription termination. Mol. Cell 3, 495–504 (1999). [PubMed: 10230402]
62. Sipos K, Szigeti R, Dong X & Turnbough CL Systematic mutagenesis of the thymidine tract of the pyrBI attenuator and its effects on intrinsic transcription termination in Escherichia coli. Mol. Microbiol 66, 127–138 (2007). [PubMed: 17725561]
63. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV & Zdobnov EM BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–2 (2015). [PubMed: 26059717]
64. Katoh K & Standley DM MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol 30, 772–80 (2013). [PubMed: 23329690]
65. Price MN, Dehal PS & Arkin AP FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490 (2010). [PubMed: 20224823]
66. Revell LJ phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol 3, 217–223 (2012).

67. Parks DH et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol 36, 996–1004 (2018). [PubMed: 30148503]

68. Bateman A et al. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169 (2017). [PubMed: 27899622]

69. Lane WJ & Darst SA Molecular Evolution of Multisubunit RNA Polymerases: Sequence Analysis. J. Mol. Biol 395, 671–685 (2010). [PubMed: 19895820]

70. Yang X et al. The structure of bacterial RNA polymerase in complex with the essential transcription elongation factor NusA. EMBO Rep. 10, 997–1002 (2009). [PubMed: 19680289]

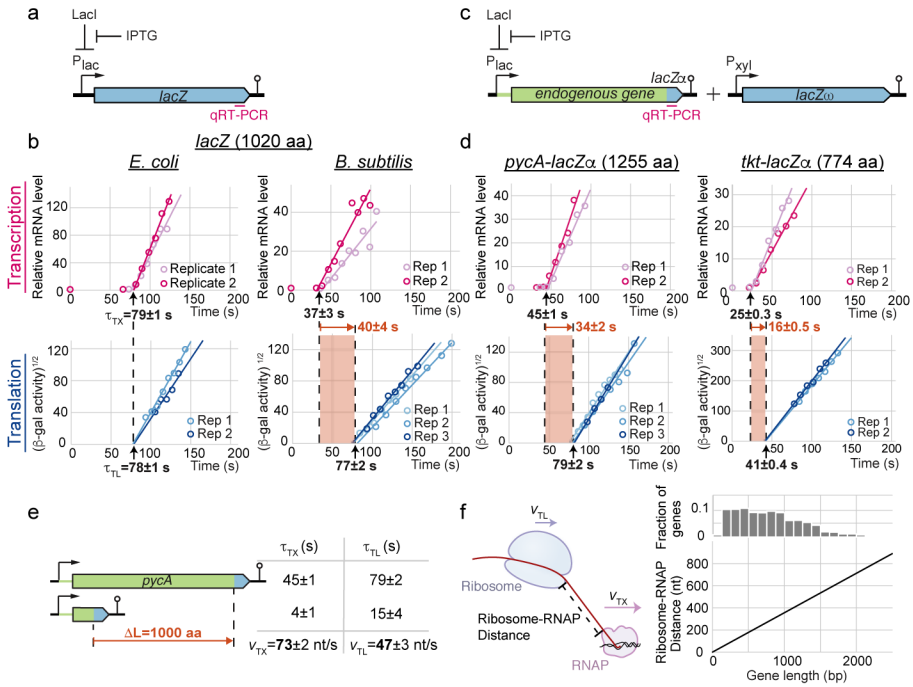


Fig. 1. Fast RNAP movement results in runaway transcription.

a, Schematic of inducible *lacZ* expression system in *B. subtilis*. Region probed by qRT-PCR is labeled in magenta. **b,** Induction time courses of full-length *lacZ* mRNA (top) and protein (bottom), measured by qRT-PCR and beta-galactosidase assays, respectively. After the first appearance times (τ_{TX} and τ_{TL}), mRNAs accumulate linearly with time whereas proteins accumulate quadratically with time. Lines indicate linear fits of transformed data after signals rise. Shaded regions indicate time difference between τ_{TX} and τ_{TL}. Uncertainties are standard error of the mean (SEM) among biological replicates (3 for *B. subtilis* beta-galactosidase assay, 2 for all others). **c,** Schematic of *lacZα* complementation reporter for endogenous genes. Endogenous 5' UTR and gene are indicated in green. P_{xyl}: xylose promoter. **d,** Same as **b**, but for endogenous genes. Translation efficiencies for *pycA* and *tkt* are 50th and 93rd percentiles among *B. subtilis* genes, respectively. Three biological replicates for *pycA-lacZα* beta-galactosidase assay, 2 for all others. **e,** Table of first appearance times for truncated *pycA* constructs. Uncertainties are standard error of the mean (SEM) among biological replicates (2). **f,** Plot showing estimated terminal ribosome-RNAP distance as a function of gene length (bottom). Elongation rates are based on **e**, and translation initiation times are assumed to be negligible. Histogram shows distribution of gene lengths in *B. subtilis* (top). See also Extended Data Fig. 1-3.

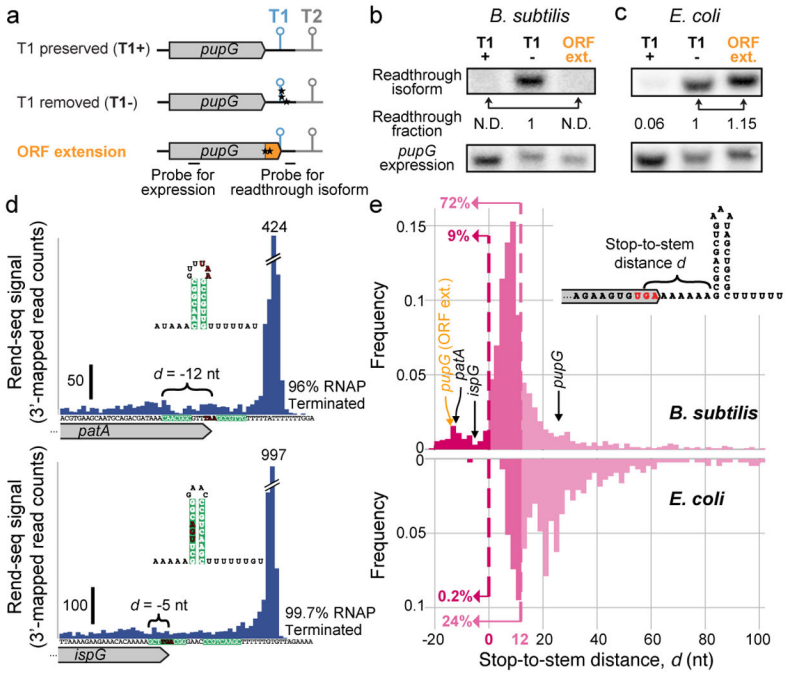


Fig. 2: Lack of translational control on transcription.

a, Schematics of ORF-extension construct and controls for *pupG* (80th percentile in translation efficiency). T1: *pupG* terminator (99.97% termination efficiency), T2: *sodA* terminator (99.9% termination efficiency). Stars indicate mutations. The stop-to-stem distances *d* for the native and extended constructs are 26 nt and minus 14 nt respectively. **b-c,** Northern blots against readthrough isoforms (top) and control for *pupG* expression (bottom) for constructs indicated in **a**. N.D.: not detected. For gel source data, see Supplementary Figure 1. Northern blotting was performed twice for *B. subtilis* (biological replicates) and once for *E. coli*. Results for both species were independently confirmed (biological replicates) by qRT-PCR (Methods). **d,** Examples of terminator stem-loops overlapping with stop codons (*patA* d=−12 nt, *ispG* d=−5 nt). Peaks in Rend-seq data show sites of termination. Terminator stems are highlighted. Stop codons are indicated in red. Translation efficiencies for *patA* and *ispG* are 63rd and 90th percentiles, respectively, in *B. subtilis*. **e,** Genome-wide distribution of stop-to-stem distances *d* (see inset) for high-confidence intrinsic terminators in *B. subtilis* (top, n=1228) and *E. coli* (bottom, n=409). ORF-overlapping terminators (d≤0) are in dark magenta, and ribosome-overlapping terminators (d≤12 nt) are in medium and dark magenta, with respective fraction of terminators indicated. See also Extended Data Fig. 5, Supplementary Data 2.

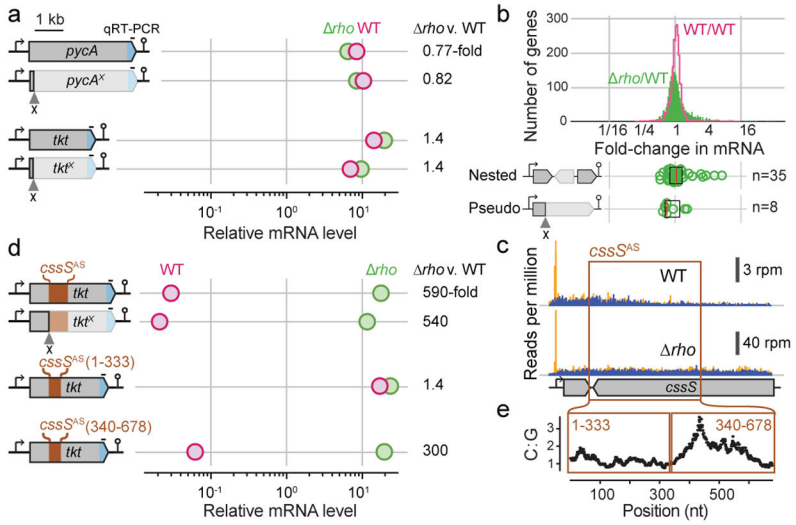


Fig. 3. Signals of Rho-dependent termination.

a, Quantification of mRNA levels with and without premature stop codons ('x'). mRNA levels are quantified by qRT-PCR for the *lacZa* region (blue) relative to *gyrA*. Comparison of mRNA levels between cells without Rho (green) and WT (magenta) are shown. **b,** Distributions of mRNA level changes between two WT replicates (magenta) and between WT and Δ*rho* (green) as measured by Rend-seq. Expression changes for asRNAs nested within operons (Extended Data Fig. 6) and pseudogenes (Extended Data Fig. 7 and 8) are indicated below. n: number of cases. Box plots are defined by median, 25th and 75th percentiles. **c,** Example of a Rho-terminated asRNA (*cssS*^{AS}). Rend-seq data in WT and Δ*rho* show regions of potential termination sites (orange: 5'-end mapped reads, blue: 3'-end mapped reads). **d,** Quantification of mRNA levels with variants of *cssS*^{AS} insertions (with 7 mutations to replace in-frame stop codons with sense codons, see Supplementary Data 1 for sequence). Relative mRNA expression measured as in **a. e,** Quantification of C-to-G ratios (number of C residues divided by number of G residues) in 100-nt moving windows of *cssS*^{AS}. See also Extended Data Fig. 6-9, Supplementary Data 3.

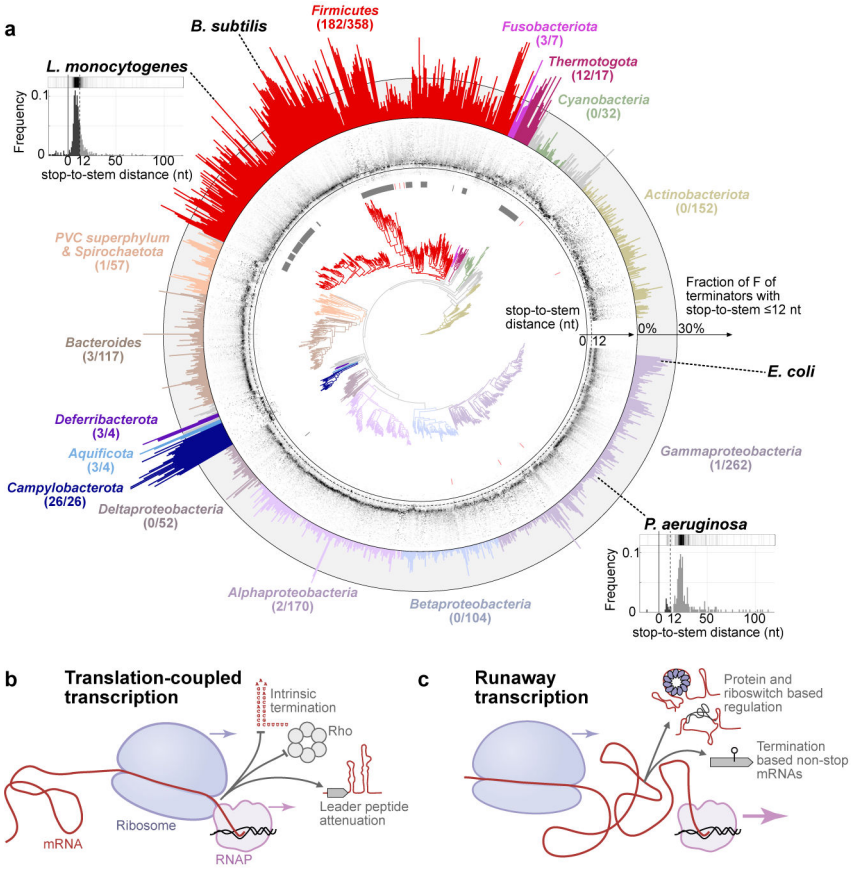


Fig. 4. Phylogenomic distribution of uncoupling.

a, Phylogenetic tree (center) is overlaid with grayscale heatmap representation of the distributions of stop-to-stem distances *d* for each species (middle ring, range in *d* shown from −20 to 120 nt). Full and dashed lines mark *d*=0 nt and *d*=12 nt respectively. The species-specific fractions F of high-confidence terminators with *d*≤12 nt is shown in the outer ring. Number of species per phylum with at least 30% of terminators with *d*≤12 nt is indicated under the phylum name. Species without Rho homologs are marked with lines next to the tree (grey: no homolog, red: partial homolog or pseudogene). The 1434 representative or reference genomes (with n≥20 identified terminators) from RefSeq are included. Tandem terminators are excluded. See Extended Data Fig. 10, Supplementary Data 4-5, Methods. Insets (*L. monocytogenes*, n=705 identified terminators, F=71.2% of terminators with *d*≤12 nt; *P. aeruginosa*, n = 216, F=6.9%) show representative examples of bioinformatically determined stop-to-stem distributions (c.f., Fig. 2e) with their heatmap representation (above) shown in middle ring. Dark and light portions of the histograms in insets highlight terminators with *d*≤12 nt and *d*>12 nt respectively. **b** and **c**, Schematics of transcription-coupled and runaway transcriptions and some of their respective functional consequences.